Improving crop productivity through data-driven optimization and hybrid deep learning-based approaches

by

Zahra Khalilzadeh

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:
Lizhi Wang, Co-major Professor
Qing Li, Co-major Professor
Guiping Hu
Sotirios Archontoulis
Brian Gelder

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2024

DEDICATION

This dissertation is dedicated to the brave women of Iran who have been fighting for their rights and freedoms for decades. I also dedicate this work to the broader Woman, Life, Freedom movement in Iran and around the world. May this dissertation contribute in some small way to the ongoing efforts to advance the cause of gender equality, human rights, and freedom for all.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I extend my sincere gratitude to those who have supported me in every step of my research journey and the writing of this thesis.

First and foremost, I express my deepest appreciation to Dr. Lizhi Wang for his unwavering guidance, patience, and support throughout this research endeavor. His invaluable insights and encouragement have been instrumental in keeping me motivated and focused, ultimately contributing to the successful completion of this work.

I am also indebted to my esteemed committee members, Dr. Guiping Hu, Dr. Qing Li, Dr. Sotirios Archontoulis, and Dr. Brian Gelder, for their valuable feedback, expertise, and contributions to refining this thesis.

Furthermore, I wish to acknowledge the significant role played by my beloved husband, Saeed. His constant love, encouragement, and patience have been my pillars of strength throughout the challenges encountered on this academic journey. My heartfelt gratitude also goes to my twin sister, Fatemeh, and her husband, Vamsi, whose unwavering support and encouragement have been a constant source of inspiration.

I am deeply grateful to my parents, Sedigheh and Mohammad Reza, for their endless love, support, and sacrifices, which have enabled me to pursue my academic aspirations. Additionally, I want to express my appreciation to my brothers, Mahdi and Mostafa, for their unwavering belief in me and their constant encouragement.

Lastly, I would like to express my gratitude to my friends, colleagues, the department faculty, and staff for enriching my experience at Iowa State University. Their friendship, encouragement, and companionship have made my time here truly memorable and fulfilling.

Each of these individuals has played a pivotal role in shaping my academic and personal growth, and for that, I am truly grateful.

ABSTRACT

With the world's population projected to approach 10 billion by 2050, coupled with the looming specter of climate change, agricultural systems are under immense pressure to enhance productivity and resilience. Climate change is expected to exert significant pressure on crop yields, exacerbating challenges already posed by increasing drought frequency, severe floods, heatwaves, and escalating pest and disease outbreaks. These environmental stressors underscore the urgency of adopting novel strategies to boost crop production and ensure food security on a global scale. Therefore, the aim of this dissertation is to address these challenges in precision agriculture. This dissertation comprises three papers that introduce novel data-driven methodologies employing optimization techniques and deep learning to address key challenges in agriculture.

In the first paper, we propose two mixed-integer linear programming (MILP) models and a heuristic algorithm to optimize planting and harvest scheduling of corn hybrids under two storage capacity cases considering both deterministic and historical growing degree unit (GDU) scenarios. Our comprehensive computational experiments and findings underscore the efficacy of our proposed methodologies. These approaches offer optimal solutions for planting and harvest scheduling, accommodating both deterministic GDU scenarios and uncertainties in historical GDU data across varying storage capacities. By ensuring consistent weekly harvest quantities below maximum capacity, our methods effectively mitigate the risks associated with inaccurate scheduling, thereby addressing logistical and productivity concerns.

In the second paper, we delve into the realm of crop yield prediction and optimal genotype selection in varying environmental conditions, with a focus on soybean hybrids. Drawing from the MLCAS2021 Crop Yield Prediction Challenge dataset, we introduce two innovative convolutional neural network (CNN) architectures: CNN-DNN and CNN-LSTM-DNN, tailored to forecast soybean yields with exceptional accuracy. To enhance the precision of yield forecasts, we propose

employing the Generalized Ensemble Method (GEM), which combines predictions from both models. Our proposed methodology exhibit RMSE reductions ranging from 5.55% to 39.88% and decreased Mean Absolute Error (MAE) ranging from 5.34% to 43.76% in comparison to baseline machine learning models, alongside higher correlation coefficients ranging from 1.1% to 10.79% when evaluated on test data. Furthermore, for optimal genotype selection, we utilize the CNN-DNN model to predict crop yields for all potential genotypes across various locations and environmental scenarios. Subsequently, we identify the top 10 genotypes with the highest yields for each location-environment combination and assess their impact on yield compared to existing genotypes. The proposed data-driven approach leads to increased average soybean yields in all states across all years.

In the third paper, we propose to use a hybrid transformer-fully connected neural networks framework to adeptly handle sequential data for corn yield prediction. Our results demonstrate the superiority of our proposed Transformer-Enhanced model compared to all other baseline machine learning models including Least Absolute Shrinkage and Selection Operator (LASSO) regression, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF), K-Nearest Neighbors (KNN), and Regression Tree (RT). Additionally, in our study, we compare transformer models with one-dimensional convolutional neural networks to assess their performance in handling sequential data. Our analysis reveal that the proposed Transformer-Enhanced model excels in handling sequential data. In this paper, we also investigate the impact of various combinations of variables on prediction errors using test data for the year 2021. Our analysis reveals the impact of dataset composition on model performance, with the variable combination that include weather and genotype data and exclude APSIM and soil datasets showing the most accurate prediction. Finally, we extend the analysis to include temporal, genomic, and geographic extrapolations to assess the robustness of the proposed Transformer-Enhanced model across different variable combinations. The results highlight that our proposed Transformer-Enhanced model effectively generalizes yield predictions to untested years, hybrids, and locations.

In conclusion, this dissertation addresses several key aspects in agriculture including corn planting and harvest scheduling under storage capacity and GDU uncertainty, soybean genotype by environment selection, and maize crop yield prediction using genotype and field data. By leveraging optimization techniques, deep learning, and data-driven approaches, we aim to pave the way for sustainable agricultural practices and ensure food security for future generations in the face of evolving environmental and demographic dynamics. The dissertation provides extensive computational experiments and results demonstrating the effectiveness of the proposed methods in addressing these critical issues in crop production.

CHAPTER 1. GENERAL INTRODUCTION

Soybean and maize are two of the most important crops globally, with significant economic and nutritional value. As the world's population continues to grow, the demand for these crops will continue to increase. Therefore, there is an urgent need to develop efficient and sustainable production systems that can maximize yields and minimize resource use. This dissertation embarks on a comprehensive exploration of critical challenges in modern agricultural practices, more specifically, this dissertation addresses the following problems:

1. Efficient crop planting and harvest scheduling
2. Optimal genotype by environment selection
3. Large scale crop yield prediction

1.1 Efficient Crop Planting and Harvest Scheduling

1.1.1 Problem Statement and Importance

The research problem we're addressing involves scheduling the planting and harvesting dates for different varieties of corn grown across two distinct sites, labeled as site 0 and site 1. Each site hosts various corn hybrids with their own specific planting windows, adding complexity to the scheduling process. Additionally, there's a strict constraint on the harvesting period, limited to a seventy-week timeframe. To tackle this problem, we're provided with extensive datasets containing historical daily growing degree units (GDU) for both sites, along with information on planting windows, required GDUs, and harvest quantities for the corn hybrids planted. Our task is to devise optimal planting and harvesting schedules for these corn hybrids under two storage capacity scenarios. The first scenario involves adhering to maximum storage capacity constraints, while the second scenario explores scheduling without such limitations to determine the minimum storage capacity required

for each site. The overarching objective is to ensure that harvesting operations do not exceed storage capacity in any given week while maintaining consistent weekly harvest quantities.

Scheduling planting and harvesting dates of corn hybrids is an important part of corn crop production. Accurate scheduling not only allows corn ears sufficient time to reach maturity but also keeps a consistent harvest amount under the storage capacity. Poor scheduling may result in inconsistent harvest quantities which can cause growers to experience logistical challenges. Moreover, it can result in having harvest quantities above the maximum capacity which might lead to either dump harvested crops or leave crops unharvested resulting in a financial loss.

1.1.2 Challenges

The first challenge that we face in this study is that the daily GDUs that are required for planting and harvest scheduling, are unknown for the scheduling year. Furthermore, each seed population possesses its unique planting window, complicating the scheduling process. Moreover, there's a stringent limitation on the harvesting period, restricted to a seventy-week timeframe. Additionally, the imperative objective is to maintain consistent weekly harvest quantities that do not exceed maximum capacity in one scenario, while also determining the lowest capacity required in scenarios where site capacities are unspecified.

1.1.3 Previous Work and Our Contributions

Several methods have been used for predicting Growing Degree Units (GDUs), ranging from the conventional linear regression model [Neild and Seeley, 1977] to the more sophisticated non-linear model [Zhou and Wang, 2018]. However, the majority of current methodologies fail to account for the influence of climate change on GDU prediction and often overlook the practicality of employing time series analysis for GDU forecasting. So, in this paper we propose to use recurrent neural networks to predict the weekly GDUs of 70 weeks and consider this as the predicted GDU scenario to solve this problem.

Growing Degree Days (GDD), also known as Growing Degree Units (GDU) or heat units, are a measure of accumulation of heat or temperature units used to estimate the plant growth stage.

GDUs are calculated based on air temperature by subtracting the base temperature from the average of the daily maximum and minimum air temperatures [The North Dakota Agricultural Weather Network (NDAWN) , 2020]:

$$\text{GDU} = (\text{Daily Maximum Air Temperature} + \text{Daily Minimum Air Temperature})/2 - \text{Base Air temperature}. \quad (1.1)$$

Base temperature is the temperature below which the crop does not grow, and it is different for different species and varieties. In the case of corn, 50°F (10°C) is often used as the base temperature. If the daily maximum temperature is above 86°F (30°C), then the daily maximum temperature is set at 86°F (30°C) as above that temperature the growth rate of corn does not significantly increase. Likewise, when the daily minimum temperature is less than 50°F (10°C), then this value is set at 50°F (10°C) [National Corn Handbook, 2020].

More recently, deep learning techniques have been utilized in many agricultural big data applications including crop yield prediction, classification of crop tolerance to heat and drought, and image-based crop yield estimation [Khaki and Wang, 2019, Khaki et al., 2020a,f,e,c, 2019, 2020d,b]. There are very few studies in the literature that use optimization models to obtain optimal planting or harvest schedule. In crop planning, Cid-Garcia et al. and Sarker et al. proposed a linear programming model to help farmers decide how to dedicate different parts of their land to different crops at different points of time to maximize their profit [Cid-Garcia et al., 2014, Sarker et al., 1997]. To the best of our knowledge our paper is the first study to propose mixed integer linear programming (MILP) models which result in optimal planting and harvesting dates for different storage capacity cases considering different GDU scenarios.

In our first study, we propose two MILP models and a heuristics algorithm to help growers and farmers schedule planting and harvesting dates of different corn populations to have consistent harvest quantities that are below the storage capacity of the site for two storage capacity cases considering a deterministic GDU scenario and multiple GDU scenarios together. To address the challenge posed by the unknown daily Growing Degree Units (GDUs) required for planting and harvest scheduling in a given year, we employ a dual approach. This involves considering both predicted GDU scenarios and historical GDU data to inform our scheduling decisions.

1.2 Optimal Genotype by Environment Selection

1.2.1 Problem Statement and Importance

The problem at hand involves determining the most suitable crop genotype to plant based on given weather conditions. Leveraging a dataset provided by Shook et al. [2021], which offers detailed genotype information on seeds, we aim to explore the feasibility of selecting genotypes based on weather variables. Employing a hybrid deep learning model, we forecast yields for all available 5,838 genotypes across diverse weather and location scenarios. Subsequently, we identify the top-performing 10 genotypes with the highest yields. After that, we calculate the average yield of these top-performing crop types for each place and weather combination. This helps us compare the suggested crop types with the ones currently being grown to see which ones give the best harvest.

Our proposed data-driven approach can be particularly valuable for selecting optimal genotypes when there are limited years of testing available. This is because the traditional approach of selecting the best genotypes based on a small number of years of field trials can be unreliable due to variations in weather and other environmental factors. By leveraging large datasets with genotype and weather information, it becomes possible to develop more accurate models that can predict the performance of different genotypes in various weather conditions. This can ultimately lead to the identification of genotypes that are both high-yielding and adaptable to different environments. Given that land for agriculture is limited, such data-driven approaches can help improve the productivity of crops per acre, as well as the quality and productivity of food crops through plant breeding.

1.2.2 Challenges

Conventionally, plant breeders rely on extensive field testing of hybrids to identify those with the highest yield potential, a process that is both time-consuming and resource-intensive. Our approach introduces a data-driven paradigm for genotype selection, wherein we use environmental data and genotype information to predict crop yields. This approach enables us to identify the

most efficient genotypes for each location and environmental condition by forecasting crop yields based on weather conditions and then selecting the optimal genotype with the highest yield.

1.2.3 Previous Work and Our Contributions

Genotype by environment interaction is a challenging factor that limits the genotype selection for increased crop yields in unseen and new environments especially with the presence of global climate change. Plant breeders typically choose hybrids based on their desired traits and characteristics, such as yield, disease resistance, and quality. They first select parent plants with desirable traits and cross them to create a new hybrid. The new hybrids are then tested in various environments to determine their performance, finally the hybrids with the highest yield are selected [Bertan et al., 2007]. However, this approach can be extremely time-consuming and tedious due to the vast number of possible parent combinations that require testing [Khaki et al., 2020a]. This highlights the importance of having a data driven approach to select genotypes with the highest performance in response to climates as well as other environmental variables using limited years of field testing per genotype. For example, Arzanipour and Olafsson [2022], suggests employing imputation methods to address the issue of incomplete data, particularly when certain crop types are not cultivated in every observed environment. This perspective views these absent data points not merely as traditional missing values but as potential opportunities for additional observations. In this study, we introduce a new deep learning framework for predicting crop yields using environmental data and genotype information. The framework is designed to identify the most efficient genotype for each location and environment, by first forecasting crop yields based on the given weather conditions in each location for all available genotypes, and then selecting the optimal genotype with the highest yield in each specific location and environmental scenario. This strategy helps in enhancing policy and agricultural decision-making, optimizing production, and guaranteeing food security. To the best of our knowledge this is the first study to use a deep learning approach for optimal genotype by environment selection.

Crop yield prediction has been more recently improved by the application of deep learning methods. Khaki and Wang [2019] utilized deep neural networks to predict corn yield for various maize hybrids using environmental data and genotype information. Their study involved designing a deep neural network model that could forecast corn yield across 2,247 locations from 2008 to 2016. With regards to the accuracy of their predictions, the model they developed outperformed others such as LASSO, shallow neural networks, and regression trees, exhibiting a Root Mean Square Error (RMSE) of 12% of the average yield when using weather data that had been predicted, and an RMSE of 11% of the average yield when using perfect weather data. Environmental data including weather and soil information and management practices were used as inputs to the CNN-RNN model developed by Khaki et al. [2020f] for corn and soybean yield prediction across the entire Corn Belt in the U.S. for the years 2016, 2017, and 2018. Their proposed CNN-RNN model outperformed other models tested including RF, deep fully connected neural networks, and LASSO, achieving a notable improvement with an RMSE of 9% and 8% for corn and soybean average yields, respectively. They also employed a guided backpropagation technique to select features and enhance the model's interpretability. Similarly, Sun et al. [2019] adopted a comparable strategy, utilizing a CNN-LSTM model to predict county-level soybean yields in the U.S. using satellite imagery, climate data, and other socioeconomic factors. Their results show that the CNN-LSTM model can capture the spatiotemporal dynamics of soybean growth and outperform other models in terms of accuracy and computational efficiency. Oikonomidis et al. [2022] utilized a publicly available soybean dataset, incorporating weather and soil parameters to develop several hybrid deep learning-based models for crop yield prediction. Comparing their models with the XGBoost algorithm, the authors found that their hybrid CNN-DNN model outperformed the other models with an impressive RMSE of 0.266, Mean Squared Error (MSE) of 0.071, and Mean Absolute Error (MAE) of 0.199. However, none of these studies have addressed the issue of determining which crop genotype to plant based on the given weather conditions.

In our second study, we design two novel convolutional neural network (CNN) architectures. The first proposed model combines CNN and fully-connected (FC) neural networks (CNN-DNN

model). The second proposed model adds a long short-term memory (LSTM) layer at the end of the CNN part for the weather variables (CNN-LSTM-DNN model). The proposed CNN-DNN model is then employed to identify the best-performing genotypes for various locations and weather conditions, making yield predictions for all potential genotypes in each specific setting. The dataset provides unique genotype information on seeds, allowing investigation of the potential of planting genotypes based on weather variables. The proposed data-driven approach can be valuable for genotype selection in scenarios with limited testing years.

1.3 Large Scale Crop Yield Prediction

1.3.1 Problem Statement and Importance

This study aims to improve crop yield prediction by using a combination of transformer and fully connected neural networks. We focus on handling sequential data, which includes things like weather patterns, crop simulation data (APSIM), and genetic information. We believe transformer models are good at capturing these complex relationships, so we design different transformer models tailored to each type of sequential data. Additionally, we include other types of data like traits and metadata in our fully connected neural network. To test our approach, we merge various datasets containing different types of information and create eight different model configurations. These configurations vary in terms of which datasets they include and exclude. Then, we compare the performance of our model with six other machine learning models and one-dimensional convolutional neural networks using data from 2021. Overall, we aim to show that our hybrid model is effective at handling sequential data and improving crop yield predictions.

Accurate prediction of crop yield holds significant advantages for global food production. It facilitates informed import and export decisions crucial for national food security, empowers farmers to make knowledgeable management choices, and enhances the efficiency of the overall food supply chain [Khaki and Wang, 2019, Jame and Cutforth, 1996, Horie et al., 1992].

1.3.2 Challenges

Predicting crop yield can be highly challenging, given the reliance on numerous intricate factors. For example, genotype information is often characterized by numerous genetic markers, each contributing minimally to the overall variance. Identifying crucial genetic markers and estimating their effects among the vast number of markers, ranging from thousands to millions, poses a considerable challenge. Furthermore, the impact of genetic markers may involve interactions with other factors, such as environmental conditions and management practices. In the third paper, we propose the utilization of transformer models to effectively handle sequential data, encompassing temporal dependencies in weather data, as well as spatial and temporal correlations in APSIM data, along with genetic linkages among adjacent genetic markers.

1.3.3 Previous Work and Our Contributions

There have been many attempts to represent the phenotype (such as yield) as an explicit function of the genotype (G), the environment (E), and their interactions ($G \times E$). Some of the earliest methods ignored the $G \times E$ interaction and just considered the additive effects of G and E, letting their interactions be treated as noise [DeLacy et al., 1996, Heslot et al., 2014]. An alternative method to study $G \times E$ is to divide the environment into some mega-environments to decrease the $G \times E$ within the mega-environment[Heslot et al., 2014]. For instance, Gauch et al. used AMMI to group environments and considered additive components for G and E as main effects and multiplicative components for $G \times E$ [Gauch Jr, 2006, Hongyu et al., 2014, Sa'diyah and Hadi, 2016]. Cooper and DeLacy used agglomerative hierarchical clustering to group environments [Cooper and DeLacy, 1994]. Some studies used factorial regression to predict $G \times E$ by identifying environmental components responsible for $G \times E$ and determining the amount of genotype sensitivity to these components [Piepho, 1998, Denis, 1988]. Crop models, sets of equations determined by a few genotypes affected by various environmental conditions, have also been used for analyzing $G \times E$ [Heslot et al., 2014, Messina et al., 2009]. However, crop models do not consider many genetic variations [Hammer et al., 2002]. Linear mixed models have been used to study $G \times E$.

Montesinos-López et al. [2016] and Montesinos-López et al. [2017] used a linear mixed model and Bayesian Poisson-lognormal method to predict multiple traits in multiple environments, explicitly considering $G \times E$ in their analysis. Lopez-Cruz et al. proposed a mixed model in which they explicitly took into account the $G \times E$. Cuevas et al. [2016] proposed a Gaussian kernel regression method incorporating a $G \times E$ mixed model and a single-environment model for prediction.

Recently, hybrid deep learning models, such as CNN-Long Short Term Memory (LSTM) model consisting of 2-Dimensional Convolutional Neural Networks (Conv2D), as well as CNN-based architectures with a 1-Dimensional convolution operation including CNN-Deep Neural Networks (DNN), CNN-LSTM, and CNN-Recurrent Neural Networks (RNN) models, have emerged to address the complexities of crop yield prediction. These challenges entail capturing both linear and non-linear relationships within diverse datasets encompassing weather, soil, climatic, and remote sensing data. These hybrid models are designed to mitigate these challenges by reducing input dimensions and extracting salient features for more accurate predictions [Oikonomidis et al., 2023]. For example, Sun et al. [2019] proposed a deep learning approach for county-level soybean yield prediction, integrating deep CNN-LSTM models. Their methodology combined crop growth and environmental variables, utilizing remote sensing data. Their study highlighted the efficacy of their proposed CNN-LSTM model over standalone CNN or LSTM models, achieving enhanced prediction accuracy for both end-of-season and in-season scenarios. Khalilzadeh et al. [2023] introduced two novel hybrid deep learning models, namely CNN-DNN and CNN-LSTM-DNN, for soybean yield prediction. These models were combined using the Generalized Ensemble Method (GEM). The one-dimensional CNN-LSTM and CNN components of their models effectively managed time dependencies in weather data, while additional factors such as genotype, maturity group, location, and year were incorporated into the fully connected (FC) part of their networks. Their findings revealed superior performance in terms of lower error rates and higher R-squared values compared to alternative machine learning models. Similarly, Srivastava et al. [2022] proposed a hybrid CNN-DNN neural network approach for winter wheat yield prediction. Like Khalilzadeh et al. [2023], their CNN component captured temporal effects of weather variables through 1-dimensional

convolution operations, while the FC network processed soil and phenology data. The integration of high-level features from CNN with FC outputs significantly enhanced yield prediction accuracy. Through evaluation against eight supervised machine learning models using root-mean-square-error(RMSE), mean absolute error(MAE), and correlation coefficients, their study demonstrated the superiority of their proposed model in predicting wheat yield. Khaki et al. [2020f] presented a novel deep learning framework combining CNNs and RNNs for accurate crop yield prediction based on environmental data and management practices. Their proposed hybrid model integrated CNNs, FC layers, and RNNs. Weather-CNN (W-CNN) and Soil-CNN (S-CNN) models captured temporal and spatial dependencies of weather and soil data, respectively. FC layers combined high-level features extracted by CNNs, while RNNs, enhanced with LSTM cells, captured temporal dynamics of crop yield trends over time. The authors applied their CNN-RNN model to predict corn and soybean yields throughout the entire Corn Belt in the United States for the years 2016-2018. Their model demonstrated significant improvements, achieving notable lower RMSE values compared to all other machine learning methods tested in their study. To the best of our knowledge, our study is the the first study to combine transformer layers with FC layers to enhance a hybrid deep learning model for crop yield prediction. By incorporating transformer layers, our proposed Transformer-Enhanced Neural Networks models effectively capture temporal dependencies in weather data, spatial and temporal correlations in environmental covariate data derived through an APSIM crop model, as well as genetic linkages among adjacent genetic markers to improve crop yield prediction accuracy.

Transformer models are a class of deep learning models that excel at processing sequential data by leveraging self-attention mechanisms to learn complex interactions among features in the data. They work by encoding input sequences and generating output sequences through attention-based mechanisms, enabling efficient learning of complex patterns in data without relying on recurrent connections. Recently, transformer based model have been used for crop yield prediction. For example, Onoufriou et al. [2023] proposed premonition network which is multi-timeline, time sequence ingesting approach based on transformer models towards processing the past, the present,

and premonitions of the future for strawberry tabletop yield forecasting. Liu et al. [2022] proposed a transformer-based model, to predict rice yield by integrating time-series satellite data, environmental variables, and rice yield records from 2001 to 2016. They showed transformer models had better performance than four other machine learning and deep learning models for end-of-season prediction. Bi et al. [2023] utilized vision transformer-based approach for soybean yield prediction using early-stage images and seed information. Lin et al. [2023] developed a novel multi-modal spatial-temporal vision transformer model for predicting crop yields at the county level across the United States, by considering the effects of short-term meteorological variations during the growing season and the long-term climate change on crops. Krishnan et al. [2024] utilized transformer models for sugarcane yield prediction.

In our third study, we introduce a novel Transformer-Enhanced Neural Networks model tailored for crop yield prediction, adept at efficiently considering diverse datasets including trait data, metadata, soil data, weather data, genotype data, and APSIM data (environmental covariate (EC) data), thereby surpassing existing methodologies in prediction accuracy. The data under consideration encompasses various types of sequential information, including temporal patterns in weather data, spatial and temporal correlations within APSIM data, and genetic associations among adjacent genetic markers. Our proposed methodology adopts a modular approach by utilizing separate TransformerEncoder models tailored to each input data category. This entails dedicated transformers designed specifically for weather, EC phenological period-soil layer, EC-phenological period, and genotype data. Such a modular design enables our model to adapt flexibly to the distinct characteristics and patterns inherent in each data category.

1.4 Dissertation Structure

This dissertation includes three papers. The first paper proposes two Mixed-Integer Linear Programming (MILP) models along with a heuristic algorithm tailored to optimize corn planting and harvest schedules while maintaining consistent weekly harvest quantities within storage capacity limits. Our approach incorporates both predicted Growing Degree Unit (GDU) scenarios

and historical data spanning ten years to ensure the optimal scheduling of planting and harvesting. The second paper introduces an ensemble framework comprising CNN-DNN and CNN-LSTM-DNN neural networks for predicting soybean yield. Additionally, the paper proposes a data-driven approach for optimal genotype selection. This approach utilizes the CNN-DNN model to predict crop yields across diverse locations and environmental scenarios, facilitating the identification of the top 10 genotypes with the highest yields for each location-environment combination. The last paper introduces a Transformer Enhanced Neural Networks framework for corn yield prediction using various combinations of sequential and non-sequential datasets.

This dissertation is organized into five chapters: Chapter 2 is dedicated to optimization models developed to determine optimal planting and harvest schedules of corn hybrids considering storage capacity and growing degree units uncertainty. Chapter 3 presents the designed ensemble framework comprising CNN-DNN and CNN-LSTM-DNN neural networks for predicting soybean yield and optimal genotype by environment selection. Crop yield prediction using Transformer-Enhanced Neural Networks and the impact of various combinations of variables on prediction errors are discussed in Chapter 4. Finally, Chapter 5 concludes the papers by discussing the contributions of their work and highlighting potential avenues for future research.

1.5 References

Arzanipour, A. and Olafsson, S. (2022). Evaluating imputation in a two-way table of means for training data construction. *SSRN*.

Bertan, I., de Carvalho, F. I., and Oliveira, A. C. d. (2007). Parental selection strategies in plant breeding programs. *Journal of crop science and biotechnology*, 10(4):211–222.

Bi, L., Wally, O., Hu, G., Tenuta, A. U., and Mueller, D. S. (2023). A transformer-based approach for early prediction of soybean yield using time-series images. *Frontiers in Plant Science*, 14:1173036.

Cid-Garcia, N. M., Bravo-Lozano, A. G., and Rios-Solis, Y. A. (2014). A crop planning and real-time irrigation method based on site-specific management zones and linear programming. *Computers and electronics in agriculture*, 107:20–28.

Cooper, M. and DeLacy, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88(5):561–572.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., Campos, G. d. l., Montesinos-López, O., and Burgueño, J. (2016). Genomic prediction of genotype× environment interaction kernel regression models. *The Plant Genome*, 9(3).

DeLacy, I., Basford, K., Cooper, M., Bull, J., McLaren, C., et al. (1996). Analysis of multi-environment trials–an historical perspective. *Plant adaptation and crop improvement*, 39124:39–124.

Denis, J. b. (1988). Two way analysis using covarites1. *Statistics*, 19(1):123–132.

Gauch Jr, H. G. (2006). Statistical analysis of yield trials by ammi and gge. *Crop science*, 46(4):1488–1500.

Hammer, G., Kropff, M., Sinclair, T., and Porter, J. (2002). Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. *European Journal of Agronomy*, 18(1-2):15–31.

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and applied genetics*, 127(2):463–480.

Hongyu, K., García-Peña, M., de Araújo, L. B., and dos Santos Dias, C. T. (2014). Statistical analysis of yield trials by ammi analysis of genotype× environment interaction. *Biometrical letters*, 51(2):89–102.

Horie, T., Yajima, M., and Nakagawa, H. (1992). Yield forecasting. *Agricultural Systems*, 40(1-3):211–236.

Jame, Y. and Cutforth, H. (1996). Crop growth models for decision support systems. *Canadian Journal of Plant Science*, 76(1):9–19.

Khaki, S., Khalilzadeh, Z., and Wang, L. (2019). Classification of crop tolerance to heat and drought—a deep convolutional neural networks approach. *Agronomy*, 9(12):833.

Khaki, S., Khalilzadeh, Z., and Wang, L. (2020a). Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach. *Plos one*, 15(5):e0233382.

Khaki, S., Pham, H., Han, Y., Kent, W., and Wang, L. (2020b). High-throughput image-based plant stand count estimation using convolutional neural networks. *arXiv preprint arXiv:2010.12552*.

Khaki, S., Pham, H., Han, Y., Kuhl, A., Kent, W., and Wang, L. (2020c). Convolutional neural networks for image-based corn kernel detection and counting. *Sensors*, 20(9):2721.

Khaki, S., Pham, H., Han, Y., Kuhl, A., Kent, W., and Wang, L. (2020d). Deepcorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation. *arXiv preprint arXiv:2007.10521*.

Khaki, S., Pham, H., and Wang, L. (2020e). Yieldnet: A convolutional neural network for simultaneous corn and soybean yield prediction based on remote sensing data. *arXiv preprint arXiv:2012.03129*.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621.

Khaki, S., Wang, L., and Archontoulis, S. V. (2020f). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750.

Khalilzadeh, Z., Kashanian, M., Khaki, S., and Wang, L. (2023). A hybrid deep learning-based approach for optimal genotype by environment selection. *arXiv preprint arXiv:2309.13021*.

Krishnan, V. G., Rao, B. S., Prasad, J. R., Pushpa, P., and Kumari, S. (2024). Sugarcane yield prediction using noa-based swin transformer model in iot smart agriculture. *Journal of Applied Biology and Biotechnology*, 12(2):239–247.

Lin, F., Crawford, S., Guillot, K., Zhang, Y., Chen, Y., Yuan, X., Chen, L., Williams, S., Minvielle, R., Xiao, X., et al. (2023). Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5774–5784.

Liu, Y., Wang, S., Chen, J., Chen, B., Wang, X., Hao, D., and Sun, L. (2022). Rice yield prediction and model interpretation based on satellite and climatic indicators using a transformer method. *Remote Sensing*, 14(19):5045.

Messina, C., Hammer, G., Dong, Z., Podlich, D., and Cooper, M. (2009). Modelling crop improvement in a g× e× m framework via gene-trait-phenotype relationships. *Crop Physiology: Interfacing with Genetic Improvement and Agronomy. The Netherlands: Elsevier*, pages 235–265.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F., Pérez-Hernández, O., Eskridge, K. M., and Rutkoski, J. (2016). A genomic bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics*, pages g3–116.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Montesinos-López, J. C., Singh, P., Juliana, P., and Salinas-Ruiz, J. (2017). A bayesian poisson-lognormal model for count data for multiple-trait multiple-environment genomic-enabled prediction. *G3: Genes, Genomes, Genetics*, pages g3–117.

National Corn Handbook (2020). https://www.extension.purdue.edu/extmedia/nch/nch-40.html.

Neild, R. E. and Seeley, M. W. (1977). Growing degree days predictions for corn and sorghum development and some applications to crop production in nebraska.

Oikonomidis, A., Catal, C., and Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, 36(1):2031822.

Oikonomidis, A., Catal, C., and Kassahun, A. (2023). Deep learning for crop yield prediction: a systematic literature review. *New Zealand Journal of Crop and Horticultural Science*, 51(1):1–26.

Onoufriou, G., Hanheide, M., and Leontidis, G. (2023). Premonition net, a multi-timeline transformer network architecture towards strawberry tabletop yield forecasting. *Computers and Electronics in Agriculture*, 208:107784.

Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics*, 97(1-2):195–201.

Sarker, R. A., Talukdar, S., and Haque, A. A. (1997). Determination of optimum crop mix for crop cultivation in bangladesh. *Applied Mathematical Modelling*, 21(10):621–632.

Sa'diyah, H. and Hadi, A. F. (2016). Ammi model for yield estimation in multi-environment trials: a comparison to blup. *Agriculture and Agricultural Science Procedia*, 9:163–169.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, 16(6):e0252402.

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., and Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*, 12(1):3215.

Sun, J., Di, L., Sun, Z., Shen, Y., and Lai, Z. (2019). County-level soybean yield prediction using deep cnn-lstm model. *Sensors*, 19(20):4363.

The North Dakota Agricultural Weather Network (NDAWN) (2020). https://ndawn.ndsu.nodak.edu/help-corn-growing-degree-days.html.

Zhou, G. and Wang, Q. (2018). A new nonlinear method for calculating growing degree days. *Scientific Reports*, 8(1):10149.

CHAPTER 2. CORN PLANTING AND HARVEST SCHEDULING UNDER STORAGE CAPACITY AND GROWING DEGREE UNITS UNCERTAINTY

Zahra Khalilzadeh[1], and Lizhi Wang[2]

[1]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[2]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

Modified from a manuscript published in *Scientific Reports*

Abstract

Planting and harvest scheduling is a crucial part of crop production due to its significant impact on other factors such as balancing the capacities for harvest, yield potential, sales price, storage, and transportation. Corn planting and harvest scheduling is challenging because corn hybrids have different planting windows, and, subsequently, inaccurate planting and harvest scheduling can result in inconsistent and unpredictable weekly harvest quantities and logistical and productivity issues. In the 2021 Syngenta Crop Challenge, participants were given several large datasets including recorded historical daily growing degree units (GDU) of two sites and provided with planting windows, required GDUs, and harvest quantities of corn hybrids planted in these two sites, and were asked to schedule planting and harvesting dates of corn hybrids under two storage capacity cases so that facilities are not over capacity in harvesting weeks and have consistent weekly harvest quantities. The research problem includes determining the planting and harvest scheduling of corn hybrids under two storage capacity cases: (1) given the maximum storage capacity, and (2) without maximum storage capacity to determine the lowest storage capacity for each site. To help improve corn planting and harvest scheduling, we propose two mixed-integer linear programming (MILP) models and a heuristic algorithm to solve this problem for both storage capacity cases. Daily GDUs are required for planting and harvest scheduling, but they are unknown at the beginning of the growing season. As such, we use

recurrent neural networks to predict the weekly GDUs of 70 weeks and consider this as the predicted GDU scenario to solve this problem. In addition, we solve this problem considering the whole given 10 historical GDU scenarios from 2010 to 2019 together for both storage capacity cases to include historical GDUs directly to our model rather than using predicted GDUs. Our extensive computational experiments and results demonstrate the effectiveness of our proposed methods, which can provide optimal planting and harvest scheduling considering deterministic GDU scenario and uncertainties in historical GDU scenarios for both storage capacity cases to provide consistent weekly harvest quantities that are below the maximum capacity.

Keywords: scheduling; mixed-integer linear programming; GDU uncertainty; recurrent neural networks; heuristic algorithm

2.1 Introduction

The advent of new farming technologies such as commercial hybrids in the 1930s preceded a widespread and rapid replacement of the once predominant open-pollinated seed varieties planted by farmers [Meyers and Rhode, 2020]. The widespread use of commercial hybrids is seen in many crops, including corn, sorghum, sugar beet, and sunflower [Wright, 1980]. Of these crops, corn is widely known as one of the world's most produced and important crops.

Scheduling planting and harvesting dates of corn hybrids is an important part of corn crop production. Accurate scheduling not only allows corn ears sufficient time to reach maturity but also keeps a consistent harvest amount under the storage capacity. Poor scheduling may result in inconsistent harvest quantities which can cause growers to experience logistical challenges. Moreover, it can result in having harvest quantities above the maximum capacity which might lead to either dump harvested crops or leave crops unharvested resulting in a financial loss. Growing Degree Days (GDD), also known as Growing Degree Units (GDU) or heat units, are a measure of accumulation of heat or temperature units used to estimate the plant growth stage. GDUs are calculated based on air temperature by subtracting the base temperature from the average of the daily maximum and minimum air temperatures [The North Dakota Agricultural Weather Network (NDAWN) , 2020]:

$$\text{GDU} = (\text{Daily Maximum Air Temperature} + \text{Daily Minimum Air Temperature})/2 - \text{Base Air temperature}. \quad (2.1)$$

Base temperature is the temperature below which the crop does not grow, and it is different for different species and varieties. In the case of corn, 50°F (10°C) is often used as the base temperature. If the daily maximum temperature is above 86°F (30°C), then the daily maximum temperature is set at 86°F (30°C) as above that temperature the growth rate of corn does not significantly increase. Likewise, when the daily minimum temperature is less than 50°F (10°C), then this value is set at 50°F (10°C) [National Corn Handbook, 2020].

Many factors affect the planting date of crops, such as weather, soil temperature, and planting resources. Traditionally, farmers and growers schedule the crops' planting dates to have a continuous harvest at the end of the growing season. Such a planting schedule has multiple benefits: (1) farmers do not have to harvest all crops at one time, and (2) farmers have some control of the harvest quantity in response to market fluctuation of crop prices to maximize their profit [Bachmann, 2008].

More recently, deep learning techniques have been utilized in many agricultural big data applications including crop yield prediction, classification of crop tolerance to heat and drought, and image-based crop yield estimation [Khaki and Wang, 2019, Khaki et al., 2020a,f,e,c, 2019, 2020d,b]. There are very few studies in the literature that use optimization models to obtain optimal planting or harvest schedule. In crop planning, Cid-Garcia et al. and Sarker et al. proposed a linear programming model to help farmers decide how to dedicate different parts of their land to different crops at different points of time to maximize their profit [Cid-Garcia et al., 2014, Sarker et al., 1997]. To the best of our knowledge our paper is the first study to propose mixed integer linear programming (MILP) models which result in optimal planting and harvesting dates for different storage capacity cases considering different GDU scenarios.

In the 2021 Syngenta Crop Challenge [Syngenta Crop Challenge, 2021], Syngenta provided real-world data and asked participants to use the data to determine optimal planting and harvest schedules of corn hybrids for two storage capacity cases. The goal is to harvest corn hybrids in a minimum number of weeks and have consistent weekly harvest quantities under the storage capacity. In this paper, we propose our approach to the 2021 Syngenta Crop Challenge. We

designed two MILP models and a heuristic algorithm for two storage capacity cases, which provide optimal planting and harvest schedules while ensuring consistent weekly harvest quantities under storage capacity. Our proposed optimization models consider both the predicted GDU scenario and all 10 historical GDU scenarios to provide optimal planting and harvest schedules. We solved the MILP models for the predicted GDU scenario using the Gurobi MILP solver and implemented the heuristic algorithm in Python for multiple GDU scenarios.

The rest of this paper is structured as follows. Section 2.2 describes the data used in this research. Section 2.3 provides a detailed description of our proposed MILP models and heuristic algorithm. Section 2.4 presents the results of our proposed models for two storage capacity cases, and two GDU scenarios including predicted GDU scenario, and all 10 GDU scenarios together. In section 2.5 we evaluate the performance of the proposed heuristic algorithm (1) by solving the corn scheduling problem considering multiple GDU scenarios using the proposed heuristic algorithm (1) and more generalized MILP model ((2.19)-(2.25)) for a small subset of populations from site 1 for storage capacity case 1 and comparing their results. Finally, the study is concluded by summarizing the key results, findings, and directions of future work in section 2.6.

2.2 Data

The dataset contained information of two separate groups of corn hybrids including 1375 and 1194 different corn seed populations planted in site 0 and site 1, respectively. The earliest and latest planting dates (planting windows) corresponding to each seed population were provided to make sure that each hybrid is planted within its planting window. Moreover, the original planting dates which are actual planting dates of the corn hybrids were provided as a benchmark. The given data also included the GDUs in Celsius for each site for each day from 2010 to 2019. Figure 2.1 shows the boxplots of the average weekly GDU of site 0 and site 1 from 2010 to 2019. We observe that site 0 has a lower median GDU than the lower quartile of site 1. As a result, crops at site 0 are expected to take longer than site 1 to accumulate necessary GDU to reach full maturity. Harvest quantities of these seed populations were provided for two storage capacity cases and their

distributions are shown in Figure 2.2 for site 0 and site 1. It can be seen that case 2 requires a higher storage capacity than case 1 for both sites.

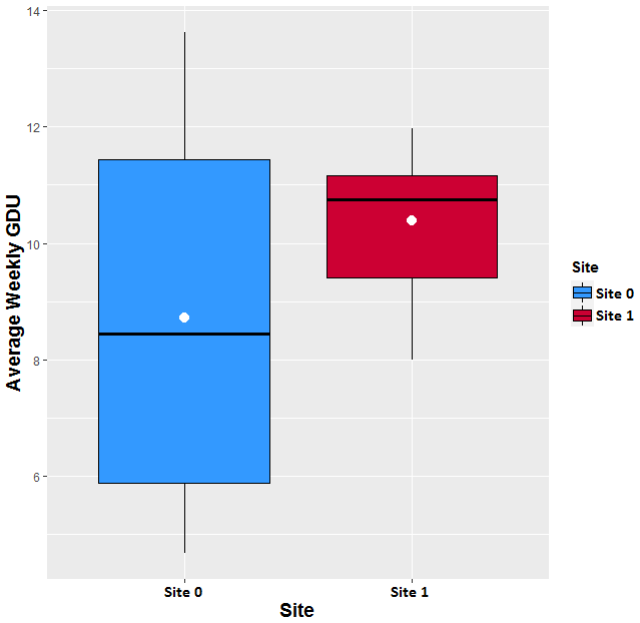


Figure 2.1: Box plot of the average weekly GDU during the last 10 years from 2010 to 2019 of each site. The white circles in each boxplots are the mean GDUs.

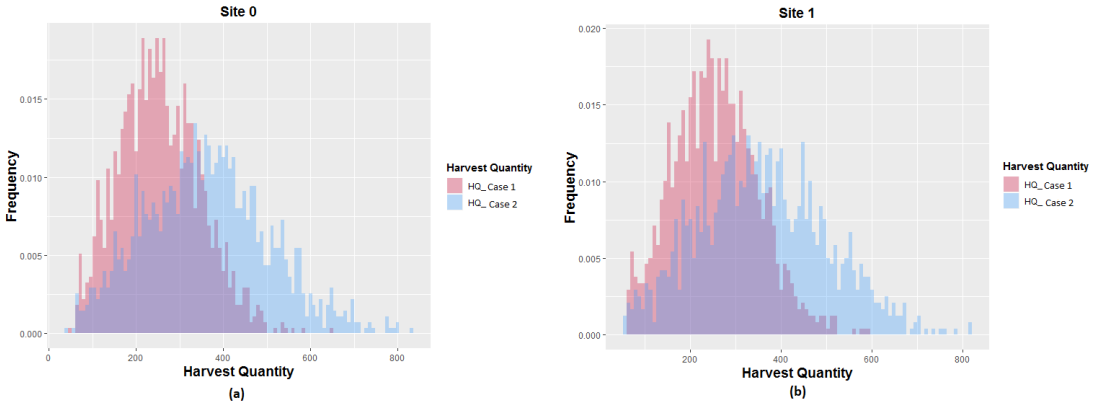


Figure 2.2: Distributions of total harvest quantity over the growing season for (a) 1375 seed populations planted in site 0, and (b) 1194 seed populations planted in site 1 for storage capacity cases 1 and 2.

2.3 Method

2.3.1 Data Preprocessing

In order to balance complexity and accuracy, we decided to model the timeline of crop growth on a weekly basis. As a result, we converted the early and late planting dates to the corresponding week numbers using the Microsoft Excel WEEKNUM function, where week 1 begins on January 1, and all subsequent weeks begin on Sundays.

Figure 2.3 shows the weekly planting window of each of 1375 and 1194 populations planted in site 0 and site 1, respectively.

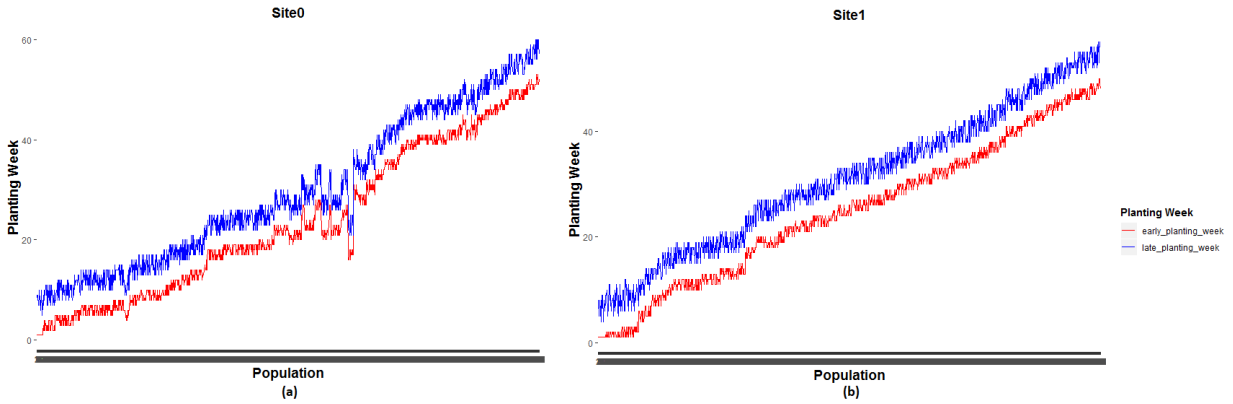


Figure 2.3: Weekly planting windows of (a) 1375 seed populations planted in site 0 and (b) 1194 seed populations planted in site 1.

2.3.2 GDU Prediction

In order to predict weekly GDUs for the 70 weeks after January 1st, 2020 using historical daily GDUs from 2010 to 2019, we designed a recurrent neural network (RNN) model, since RNN models can capture temporal dependencies. Long short-term memory (LSTM) Hochreiter and Schmidhuber [1997] model is a type of recurrent neural network, which can capture long-term time dependencies in the the data without having problems such as vanishing gradients. An LSTM unit is usually composed of cell, an input gate, an output gate and a forget gate which control the flow

of information through time steps. We considered the following three versions of LSTM models to determine the best model structure for the two sites.

Vanilla RNN: Vanilla RNN is a basic version of recurrent neural network where they use their internal hidden state units (memory) to capture the temporal effects of data.

Bidirectional LSTM: Bidirectional LSTMs are an extension of traditional LSTMs where they process the temporal information in both directions backwards or forward.

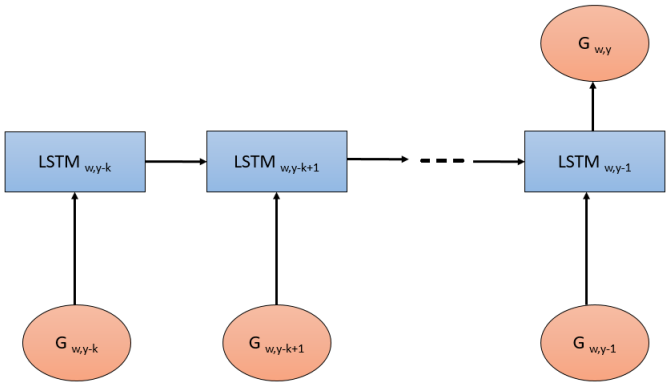
Stacked LSTM: We used two hidden layers each with 50 LSTM units each using ReLU activation function. The model used Adam optimizer and stochastic gradient descent to optimize the mean squared error (MSE) loss function.

These three LSTM models were trained using all weekly GDUs from 2010 to 2019. Let G_y^w denote the GDU of week w in year y with $\forall w \in \{1, \ldots, 52\}$ and $\forall y \in \{2010, \ldots, 2019\}$. The LSTM model explains the GDU variable G_y^w as a response of k previous years in the same week: $\{G_{y-k}^w, G_{y-k+1}^w, G_{y-k+2}^w, G_{y-k+3}^w, ..., G_{y-1}^w\}$. We considered three periodic lags including 3, 4, and 5 and found that 3 years to yield the best results. As such, we used the 312 weekly GDUs from 2010 to 2018 as the training data and the 52 weekly GDUs in 2019 as the validation data to compare the aforementioned LSTM models. We used root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient as comparison criteria. Results are shown in Table 2.1, which suggest that Stacked LSTM and Bidirectional LSTM with 3 years periodic lag had the best performance for site 0 and site 1, respectively.

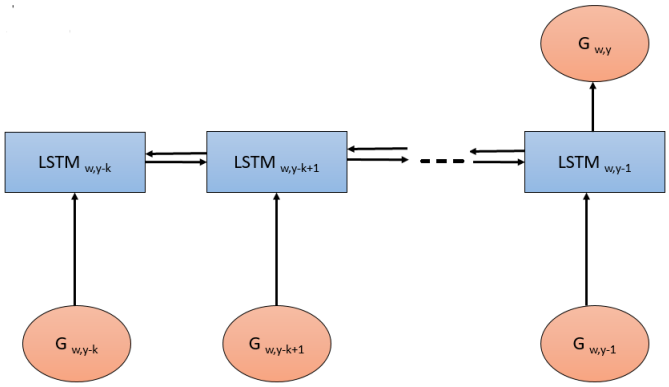
Table 2.1: Comparison of predictive performances of three LSTM models for two sites.

Site	Method	Lag	RMSE	MAE	Correlation Coefficient(%)	Number of Samples
Site0	Vanilla LSTM	3	7.21	6.06	0.98	312
		4	8.30	7.25	0.97	260
		5	7.92	6.59	0.98	208
	Bidirectional LSTM	3	9.76	6.82	0.93	312
		4	8.01	6.97	0.98	260
		5	9.85	8.73	0.98	208
	Stacked LSTM	**3**	**7.07**	**5.91**	**0.98**	**312**
		4	8.45	7.29	0.97	260
		5	7.77	6.34	0.97	208
Site1	Vanilla LSTM	3	5.43	4.24	0.83	312
		4	6.37	4.72	0.81	260
		5	6.47	4.46	0.84	208
	Bidirectional LSTM	**3**	**5.19**	**3.93**	**0.83**	**312**
		4	6.54	4.57	0.80	260
		5	5.63	4.30	0.83	208
	Stacked LSTM	3	6.24	5.18	0.83	312
		4	5.49	4.02	0.82	260
		5	6.77	5.56	0.83	208

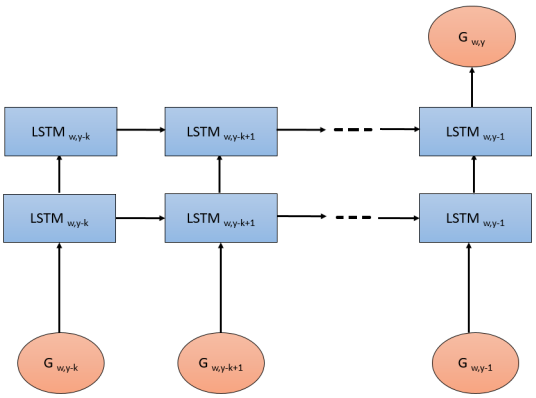
The Stacked and Bidirectional LSTM networks were then trained again to predict G_y^w for the year 2020 using weekly GDU data from 2017 to 2019. We then we used the predicted 2020 data and historical data of 2018 and 2019 to predict weekly GDU in 2021 for the first 20 weeks. The structures of LSTM networks are illustrated in Figure 2.4.



(a)



(b)



(c)

Figure 2.4: Three LSTM models for GDU prediction with a k-year lag. Subfigures (a), (b), and (c) are for vanilla, bidirectional, and stacked LSTMs, respectively.

2.3.3 Optimization Models

In this section we propose four optimization models for scheduling planting and harvesting dates of seed populations. The first two models are for case 1, in which storage capacities for sites 0 and 1 are given as 7000 and 6000 ears, respectively. The first model is a deterministic one, considering a single GDU scenario with predicted weekly GDU values, and the second model is a stochastic one, considering ten historical years of weekly GDU data as ten scenarios. The third and forth models are, respectively, deterministic and stochastic models for case 2, in which storage capacities for the two sites are decision variables.

2.3.3.1 Deterministic Model for Case 1

In this model, predicted weekly GDUs from section 2.3.2 were used. This model consists of the following decision variables:

$$t_{ij}^{\mathrm{p}} = \begin{cases} 1, & \text{if population } i \text{ is planted in week } j \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

$$t_{ij}^{\mathrm{h}} = \begin{cases} 1, & \text{if population } i \text{ is harvested in week } j \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

$$w_j = \begin{cases} 1, & \text{if any population is harvested in week } j \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

Parameters of this model include:

- C: storage capacity of a site
- GDU_j: cumulative weekly GDU from week 1 to week j
- N: total number of populations
- T: number of weeks after January 1st of planning year. In the 2021 Syngenta crop challenge, $T = 70$.

- HQ_i: harvest quantity (number of ears) of population i
- $T^{p_i} = \{\ [l_i, u_i]\ \}$: the planting window for the population i, where l_i and u_i are the corresponding earliest and latest planting dates, respectively.
- $G_i^{\min}$: number of GDUs needed by population i before harvesting

We formulate our optimization model as the following:

$$\min_{t_{ij}^{\mathrm{p}}, t_{ij}^{\mathrm{h}}, w_j} \quad \sum_{j=1}^{T} |w_j C - \sum_{i=1}^{N} HQ_i t_{ij}^{\mathrm{h}}| \tag{2.5}$$

$$s.t. \quad \sum_{j \in T^{p_i}} t_{ij}^{\mathrm{p}} = 1 \qquad \forall i \in \{1, ..., N\} \tag{2.6}$$

$$\sum_{j \notin T^{p_i}} t_{ij}^{\mathrm{p}} = 0 \qquad \forall i \in \{1, ..., N\} \tag{2.7}$$

$$\sum_{j=1}^{T} t_{ij}^{\mathrm{h}} = 1 \qquad \forall i \in \{1, ..., N\} \tag{2.8}$$

$$\sum_{j=1}^{T} (t_{ij}^{\mathrm{h}}\ GDU_j - t_{ij}^{\mathrm{p}}\ GDU_j)\ \geq\ G_i^{min} \qquad \forall i \in \{1, ..., N\} \tag{2.9}$$

$$\sum_{j=1}^{T} (t_{ij}^{\mathrm{h}}\ GDU_{j-1} - t_{ij}^{\mathrm{p}}\ GDU_j) \leq\ G_i^{min} - 1 \quad \forall i \in \{1, ..., N\} \tag{2.10}$$

$$N\ w_j\ \geq\ \sum_{i=1}^{N} t_{ij}^{\mathrm{h}} \qquad \forall j \in \{1, ..., T\} \tag{2.11}$$

$$t_{ij}^{\mathrm{p}},\ t_{ij}^{\mathrm{h}},\ w_j \in \{0, 1\} \qquad \forall i \in \{1, ..., N\},\ \forall j \in \{1, ..., T\} \tag{2.12}$$

Here, the objective (2.5) is to minimize the difference between the weekly harvest quantity and the capacity for each harvesting week while using the minimum number of weeks for harvesting. We adopt the absolute value function in our objective function because it can be easily linearized and is computationally more tractable [Ferguson, 2000]. Constraints (2.6) and (2.7) make sure that each population is planted within its corresponding planting window. Constraint (2.8) means that each population can only be harvested in one week. Constraints (2.9) and (2.10) enforce the model to harvest populations as soon as they accumulate their required GDU. Constraint (2.11) requires that $w_j = 1$ when any population is harvested in week j. Finally, Constraint (2.12) defines all decision variables as binary.

Due to having the absolute value function inside our objective, the above-mentioned model is a nonlinear optimization problem which is hard to solve. As a result, we reformulate our model into

an equivalent mixed integer linear programming (MILP) by introducing a set of new variables which is as follows [Ferguson, 2000]:

$$\min_{t_{ij}^{\mathrm{p}},t_{ij}^{\mathrm{h}},w_j,e_j^+,e_j^-} \quad \sum_{j=1}^{T}(e_j^+ + e_j^-) \tag{2.13}$$

$$w_j C - \sum_{i=1}^{N} HQ_i\, t_{ij}^{\mathrm{h}} = e_j^+ - e_j^- \qquad \forall j \in \{1,...,T\} \tag{2.14}$$

$$\text{Constraints } (2.6) - (2.11) \tag{2.15}$$

$$t_{ij}^{\mathrm{p}},\ t_{ij}^{\mathrm{h}},\ w_j \in \{0,1\},\ \ e_j^+, e_j^- \geq 0 \qquad \forall i \in \{1,...,N\},\ \forall j \in \{1,...,T\} \tag{2.16}$$

Here, the objective (2.13) is to minimize the positive and negative errors between the storage capacity and the sum of harvested quantities for each harvesting week which is equivalent to objective (2.5). In constraint (2.14) we define the two error terms e_j^+ and e_j^-. Because in this problem the goal is to have weekly harvest quantities under the capacity, we put e_j^- zero in the optimal solution. As a result, the model will be improved by reducing the positive error. Constraint (2.16) indicates the appropriate types of the decision variables.

2.3.3.2 Stochastic Model for Case 1

Considering multiple GDU scenarios during the last ten years from 2010 to 2019 together, we now propose a more generalized optimization model to find the optimal planting dates of all populations for all 10 historical GDU scenarios so that based on the weekly GDU values of each of ten GDU scenarios the corn populations will be harvested on different dates. Then the maximum weekly harvest quantity among all 10 GDU scenarios will be consistent and below the capacity of

the site. The decision variables of the generalized optimization model are:

$$t_{ij}^{\mathrm{p}} = \begin{cases} 1, & \text{if population } i \text{ is planted in week } j \\ 0, & \text{otherwise} \end{cases} \tag{2.17}$$

$$t_{ij}^{\mathrm{h}k} = \begin{cases} 1, & \text{if population } i \text{ is harvested in week } j \text{ with GDU of year } k \\ 0, & \text{otherwise} \end{cases} \tag{2.18}$$

This optimization model consists of the following notations:

- k: index of the year corresponding to different GDU scenarios from 2010 to 2019 of each site.
- C: the storage capacity of each site
- A_{jk}: the harvest quantity of week j of GDU scenario of year k.
- K: number of years which is 10.
- $A_{j\max}$: the maximum weekly harvest quantity among all GDU scenarios of K years.
- GDU_j^k: the cumulative weekly GDU of year k which is accumulated till week j
- N: the total number of populations
- T: number of weeks after Jan 1st of planning year (it is 70 weeks for the 2021 Syngenta crop challenge.)
- HQ_i: number of ears (harvest quantity) of case 1 produced by population i
- $T^{p_i} = \{\, [l_i, u_i] \,\}$: the planting window for the population i, where l_i and u_i are the corresponding earliest and latest planting dates, respectively.
- $G_i^{\min}$: number of growing degree units needed before harvesting population i (required GDUs for population i)

As such, we define our more generalized optimization model to consider all GDU scenarios and find one set of optimal planting dates which works for all GDU scenarios:

$$\min_{t_{ij}^{\mathrm{p}}, t_{ij}^{\mathrm{h}k}, w_j^k} \sum_{j=1}^{T} |C - A_{j\max}| \tag{2.19}$$

$$A_{jk} \;=\; \sum_{i=1}^{N} HQ_i \, t_{ij}^{\mathrm{h}k} \qquad \forall j \in \{1, ..., T\} \qquad \forall k \in \{1, ..., K\} \tag{2.20}$$

$$A_{j\max} \;=\; max_k(A_{jk}) \qquad \forall j \in \{1, ..., T\} \tag{2.21}$$

$$s.t. \sum_{j \in T^{p_i}} t_{ij}^{\mathrm{p}} = 1 \qquad \forall i \in \{1, ..., N\} \tag{2.22}$$

$$\sum_{j \notin T^{p_i}} t_{ij}^{\mathrm{p}} = 0 \qquad \forall i \in \{1, ..., N\} \tag{2.23}$$

$$\sum_{j=1}^{T} t_{ij}^{\mathrm{h}k} = 1 \qquad \forall i \in \{1, ..., N\} \qquad \forall k \in \{1, ..., K\} \tag{2.24}$$

$$\sum_{j=1}^{T} t_{ij}^{\mathrm{h}k} \, GDU_j^k - t_{ij}^{\mathrm{p}} \, GDU_j^k \;=\; G_i^{min} \qquad \forall i \in \{1, ..., N\} \qquad \forall k \in \{1, ..., K\} \tag{2.25}$$

$$t_{ij}^{\mathrm{p}},\, t_{ij}^{\mathrm{h}k} \in \{0, 1\} \qquad \forall i \in \{1, ..., N\},\, \forall j \in \{1, ..., T\} \qquad \forall k \in \{1, ..., K\} \tag{2.26}$$

The objective function (2.19) is to minimize the sum of differences between maximum weekly harvest quantities among 10 GDU scenarios and the capacity of the site. The key to solving model

(2.19) is t_{ij}^{p} that represents the planting dates of the populations. Once this variable is revealed, then given each GDU scenario's weekly GDU values the harvesting weeks will be known. It is because the populations should be harvested as soon as they reach their required GDUs. So, given the planting schedule of corn populations because of the different weekly GDUs for each GDU scenario the harvesting week of populations and subsequently weekly harvest quantities would be different for each GDU scenario.

2.3.3.3 Heuristic Algorithm

Unlike the optimization model (2.13)-(2.16) proposed in 2.3.3.1 which can be solved using the existing branch-and-bound algorithms Lawler and Wood [1966], solving the optimization model (2.19)-(2.26) using existing algorithms and solvers is extremely time-consuming due to numerous number of variables. Therefore, this section presents a metaheuristic algorithm, simulated annealing (SA) [Van Laarhoven and Aarts, 1987], for solving the corn scheduling problem considering multiple GDU scenarios together (2.19)-(2.26), which is computationally tractable and searches for a local optimal solution to model (2.19)-(2.26). Simulated annealing is a probabilistic optimization method designed for finding the global minimum of a cost function that may possess several local minima. SA algorithm is based on the emulation of physical annealing process where a heated solid is cooled down to reach a minimum energy configuration [Bertsimas et al., 1993]. Our goal is to find an optimal planting date, t^{p^*}, considering GDU scenarios of 10 different years which minimizes the objective function defined in Equation (2.19). In our problem definition, we no longer need to include the harvest dates variables in our model because of the assumption of harvesting corn populations as soon as they accumulate their respective minimum required GDUs. Our heuristic algorithm starts with a random solution (x_0) and initial temperature (T_0), and continues until a maximum of k_{max} iterations. Steps of the heuristic algorithm used in this study are as follows:

Algorithm 1: Heuristic Algorithm

Result: optimal planting date, t^{p^*}, considering 10 GDU scenarios together from different years

start with a random solution (t_0^{p}) and initial temperature (T_0) ;

Let $T = T_0$ and $t^{\mathrm{p}} = t_0^{\mathrm{p}}$

for $k = 0$ *through* k_{max} **do**

  $T \leftarrow \Lambda(T, k, k_{max})$

  Create a random neighbor, $t_{\mathrm{new}}^{\mathrm{p}} \leftarrow \Omega(t^{\mathrm{p}})$

  if $F_{prob}(E(t^p)\ ,\ E(t_{new})\ ,\ T\ ) \geq$ *random(0,1)* **then**

    $t^{\mathrm{p}} \leftarrow t_{new}^{\mathrm{p}}$;

  end

end

Here, $\Lambda(T, k, k_{max})$, $\Omega(t^{\mathrm{p}})$, and $F_{prob}(E(t^{\mathrm{p}})\ ,\ E(t_{\mathrm{new}})\ ,\ T\ )$ are temperature, create-neighbor, and energy functions, respectively. We used the following temperature function to decay the initial temperature (T_0) during the heuristic algorithm process:
$\Lambda(T, k, k_{max}) = T_0\ \alpha^k,$ where $k \leq k_{max},$ and $0 \leq \alpha \leq 1$. Higher temperature allows more exploration of the solution space at the beginning of the optimization process. We set max iteration, decay rate (α), and initial temperature to be 700, 0.995, and 30000, respectively. We designed a create-neighbor function, $\Omega(t^{\mathrm{p}})$, which creates a new solution based on the previous solution which considers both the exploitation of current solution and exploration of solution space. As such, $\Omega(t^{\mathrm{p}})$ first finds the weeks which have the maximum and minimum harvest quantities across all GDU scenarios. For the populations at the maximum harvest quantity week, we select three populations with minimum harvest quantities and randomly move either forward or backward their respective planting dates one or two weeks. For the week with minimum harvest quantity, we try to move populations from closest weeks with the larger harvest quantities to the week with minimum harvest quantity. We used the following energy function to compute the probability of the acceptance of the current solution: $F_{prob}(E(t^{\mathrm{p}})\ ,\ E(t_{\mathrm{new}})\ ,\ T\ ) = e^{-(E(t^{\mathrm{p}})-E(t_{\mathrm{new}}))/T},$ where

$E(t^{\mathrm{p}})$, $E(t_{\mathrm{new}})$, and T are the cost of best solution so far, the cost of new solution, and the current temperature, respectively.

2.3.3.4 Deterministic Model for Case 2

This subsection illustrates our proposed optimization model for case 2 where there is not a predefined capacity, and the goal is to determine planting and harvesting dates of each population (during T weeks) and also the lowest capacity required for each site. The optimization model for case 2 has the same decision variables and constraints with the optimization model proposed for case 1, but the objective function is changed to achieve the goal of case 2. Here θ_w is a coefficient for the number of harvesting weeks, and it is set to be 1. The optimization model for case 2 is as follows:

$$\min \quad \max \ \{\sum_{i=1}^{N} HQ_i\, t_{i1}^{\mathrm{h}}, \sum_{i=1}^{N} HQ_i\, t_{i2}^{\mathrm{h}}, ..., \sum_{i=1}^{N} HQ_i\, t_{iT}^{\mathrm{h}}\} + \theta_w \sum_{j=1}^{T} w_j \tag{2.27}$$

$$\text{Constraints } (2.6) - (2.10), (2.14) \tag{2.28}$$

$$t_{ij}^{\mathrm{p}},\ t_{ij}^{\mathrm{h}},\ w_j \in \{0,1\},\ \ e_j^{+}, e_j^{-} \geq 0 \quad \forall i \in \{1,...,N\},\ \forall j \in \{1,...,T\} \tag{2.29}$$

This is a Minimax Linear Programming Problem (MLPP), and the objective (2.27) is to minimize the maximum amount of weekly harvest quantities which simultaneously minimizes the amount of weekly harvest quantities for all weeks while using the minimum number of harvesting weeks. Since the objective (2.27) is nonlinear, we reformulate our model into an equivalent mixed integer linear programming (MILP) by introducing a new variable denoted by z [Ahuja, 1985]:

$$\min \quad z + \theta_w \sum_{j=1}^{T} w_j \tag{2.30}$$

$$z \geq \sum_{i=1}^{N} HQ_i \, t_{ij}^{\mathrm{h}} \quad \forall j \in \{1, ..., T\} \tag{2.31}$$

$$\text{Constraints} \quad (2.6) - (2.10), (2.14) \tag{2.32}$$

$$t_{ij}^{\mathrm{p}},\ t_{ij}^{\mathrm{h}},\ w_j \in \{0, 1\},\ \ e_j^{+}, e_j^{-}, z \geq 0 \quad \forall i \in \{1, ..., N\},\ \forall j \in \{1, ..., T\} \tag{2.33}$$

Here, the objective is to minimize the summation of the maximum value of the weekly harvest quantities and the total number of harvesting weeks. Constraint (2.31) ensures that the maximum value of the weekly harvest quantities is always greater than or equal to the amount of harvest quantity of each week.

2.3.3.5 Stochastic Model for Case 2

As it was mentioned previously, in case 2, there is not a predefined capacity for each site and we need to determine the lowest capacity required. This subsection presents a more generalized optimization model which considers all ten GDU scenarios together to determine the optimal planting dates of all populations. The optimal planting dates of all populations will result in different harvesting dates based on the weekly GDU values of each of ten GDU scenarios. The goal here is to determine the lowest required capacity in a way that the maximum weekly harvest quantities among all 10 GDU scenarios will be consistent.

To this end, we define a new loss function which computes the sum of the absolute differences for each weekly harvest quantity which is the maximum weekly harvest quantity among all GDU scenarios and its neighboring weekly harvest quantities. Minimizing this loss function ensures consistent harvest quantities with lowest possible storage capacity. We use the same heuristic algorithm described in section 2.3.3.3.

2.4 Results

In this section, we present the quantitative results of our optimization models for case 1 and case 2 considering predicted GDU scenario, and all 10 GDU scenarios together. In the 2021 Syngenta crop challenge, participants were asked to schedule the planting date of each seed population sometimes within the given planting window for each seed population and the harvesting dates during 70 weeks after January 1st 2020. As it was discussed in section 2.3.2 the daily GDUs of these 70 weeks were unknown a priori and as a part of the challenge we made use of historical daily GDUs from 2010 to 2019 provided for each site to estimate GDUs of these 70 weeks for each site which is the predicted GDU scenario. We also solved the optimization models considering all 10 historical GDUs together.

We implemented our MILP models (2.13)-(2.16) and (2.30)-(2.33) for case 1 and case 2 respectively considering predicted GDU scenario in MATLAB R2018a and solved with MILP commercial solver Gurobi Optimizer. The heuristic algorithm considering multiple GDU scenarios that include all 10 historical GDUs was implemented in Python for both cases 1 and 2. A summary of results from the deterministic and stochastic models for storage capacity cases 1 and 2 for both sites is provided in Table 2.2.

2.4.1 Results of the Deterministic Model for Case 1

The MILP model for case 1 (2.13)-(2.16) considering predicted GDU scenario which was calculated in section 2.3.2 was run in MATLAB for each site. Input variables for the optimization model for each site include: the storage capacity (C) of 7000 ears for site 0 and 6000 ears for site 1, total number of populations (N) of 1375 planted in site 0 and 1194 planted in site 1, total number of ears produced by each seed population (HQ_i) planted in site 0 and site 1 given for case 1, and required GDUs of each seed population ($G_i^{\min}$) planted in site 0 and site 1. The cumulative weekly GDUs (GDU_j) of each site are calculated using the predicted weekly GDU from section 2.3.2 for each site. The number of weeks after Jan 1st of planning year (T) is equal to 70 for both sites.

Finally, the planting windows of seed populations are given to the model so that the planting date for each seed population falls into its corresponding planting window (T^{p_i}).

Table 2.2: Summary of results from deterministic and stochastic models for storage capacity cases 1 and 2 for both sites.

Site No.-Capacity Case	GDU Scenario	Optimization Model	Difference between Harvest Quantities and Capacity for Original Planting Dates	Difference between Harvest Quantities and Capacity for Optimal Planting Dates	Lowest Storage Capacity for Original Planting Dates	Lowest Storage Capacity for Optimal Planting Dates
Site 0-Case 1	Det.	MILP	223,373	6,793	–	–
Site 1-Case 1	Det.	MILP	116,153	4,661	–	–
Site 0-Case 1	Stoch.	Both MILP and heuristic are infeasible	–	–	–	–
Site 1-Case 1	Stoch.	Heuristic	172,851	87,421	–	–
Site 0-Case 2	Det.	MILP	–	–	37,247	10,795
Site 1-Case 2	Det.	MILP	–	–	16,220	8,108
Site 0-Case 2	Stoch.	Both MILP and heuristic are infeasible	–	–	–	–
Site 1-Case 2	Stoch.	Heuristic	–	–	21,811	11,192

Weekly harvest quantities considering the predicted GDU scenario for both optimal planting dates and original planting dates are shown in Figures 2.5 and 2.6 for site 0 and site 1 respectively. Optimal planting dates are the optimal planting dates resulted from our proposed MILP model for case 1 (2.13)-(2.16) considering the predicted GDU scenario, and original planting dates are the actual planting dates of the populations which were given by the Syngenta Crop Challenge. These figures suggest that the proposed MILP model (2.13)-(2.16) was able to schedule the planting and harvesting dates of the whole 1375 and 1194 populations planted in site 0 and site 1 respectively

with different planting windows, required GDUs, and harvest quantities in a way that resulted in consistent weekly harvest quantities that are below the storage capacities.

The absolute maximum and absolute median difference between the weekly harvest quantity and the capacity among all harvesting weeks for case 1 considering the predicted GDU scenario for optimal and original planting dates for site 0 and site 1 are shown in Table 2.3. Moreover, run-time, and values of both objectives including number of harvesting weeks and sum of absolute differences between the capacity and weekly harvest quantities resulted from the proposed model (2.13)-(2.16) are presented in Table 2.4 for both sites.

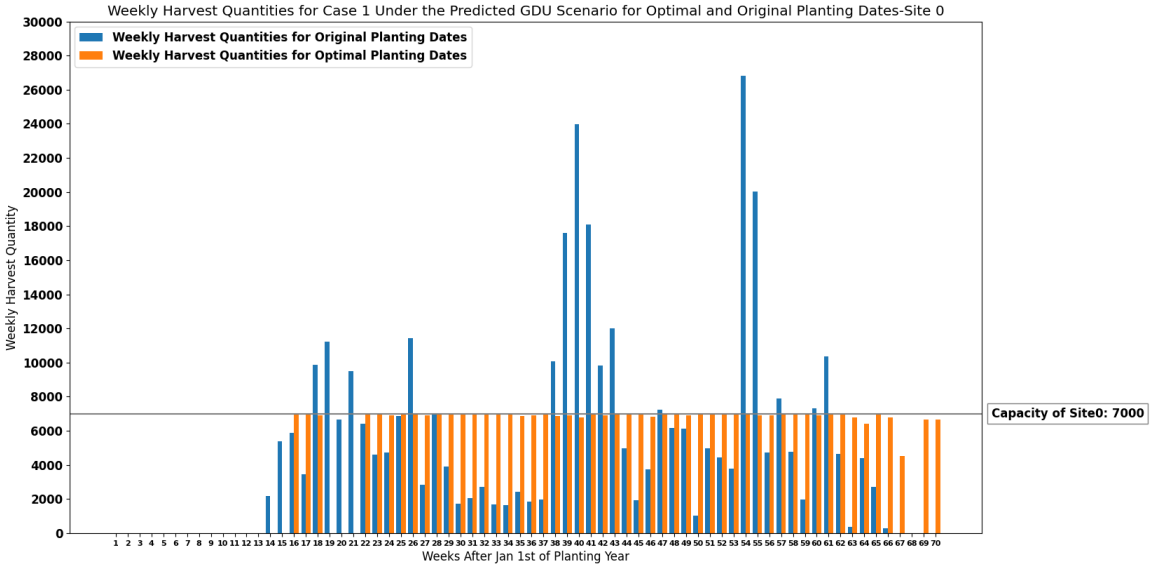


Figure 2.5: Weekly harvest quantities of site 0 with a capacity of 7000 ears for case 1 under the predicted GDU scenario for optimal and original planting dates.

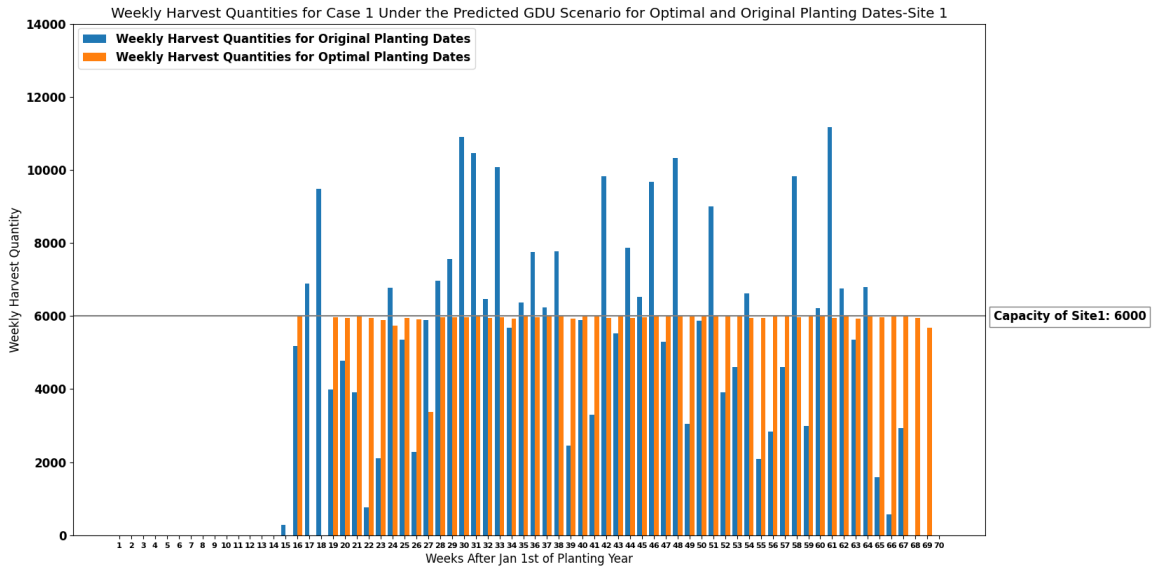


Figure 2.6: Weekly harvest quantities of site 1 with a capacity of 6000 ears for case 1 under the predicted GDU scenario for optimal and original planting dates.

Table 2.3: The absolute median difference between the weekly harvest quantity and the capacity (Median Dif) and the absolute maximum difference between the weekly harvest quantity and the capacity (Maximum Dif) among all harvesting weeks for site 0 and site 1 for case 1 under the predicted GDU scenario.

Site	Optimal Planting		Original Planting	
	Median Dif	Maximum Dif	Median Dif	Maximum Dif
0	57	2471	2220	6732
1	25	2633	118	5718

Table 2.4: Run-time, and values of both objectives including number of harvesting weeks and sum of absolute differences between the capacity and weekly harvest quantities resulted from the proposed MILP model for case 1 under the predicted GDU scenario for site 0 and site 1.

Site	Run-time (seconds)	Number of harvesting weeks	Sum of absolute differences
0	3191	51	6793
1	65	52	4661

Figures 2.7 and 2.8 present optimal planting weeks suggested by our proposed MILP model (2.13)-(2.16) for case 1 under the predicted GDU scenario along side with the early and late planting weeks for the the whole 1375 seed populations planted in site 0 and 1194 seed populations planted in site 1 respectively. These plots show that the optimal planting date for each seed population falls into its corresponding planting window.

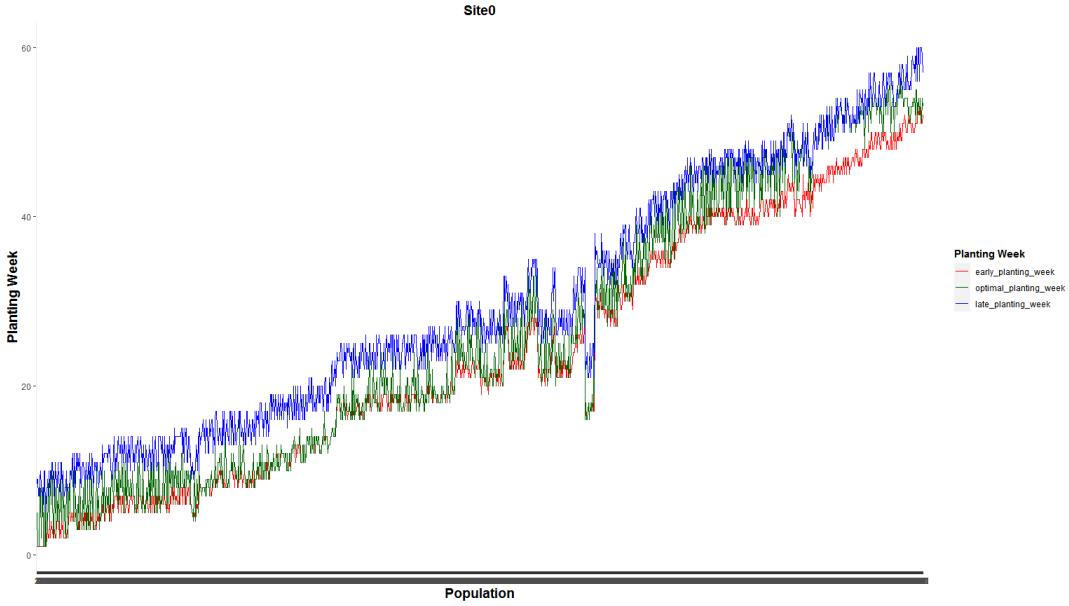


Figure 2.7: Optimal planting weeks along side with the early and late planting weeks for the the whole 1375 seed populations planted in site 0 for case 1 under the predicted GDU scenario.

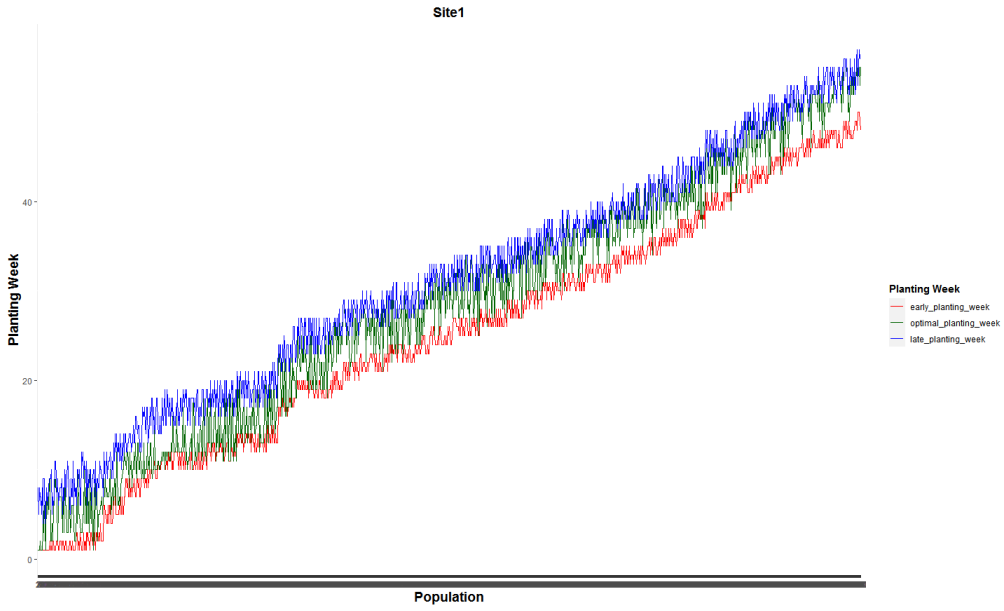


Figure 2.8: Optimal planting weeks along side with the early and late planting weeks for the whole 1194 seed populations planted in site 1 for case 1 under the predicted GDU scenario.

2.4.2 Results of the Stochastic Model for Case 1

In this subsection we present the results of our proposed heuristic algorithm for case 1. As it was discussed in section 4.3 there is a lower amount of heat available for the growth of crops in site 0. As a result the growing degree units required in order for the corn population to achieve maturity accumulate slower in site 0 and the seed populations may need more than 70 weeks to be ready to be harvested. So, for site 0 all 10 GDU scenarios do not enable us to harvest the whole 1375 seed populations planted in site 0 in 70 weeks. So, we only solved our heuristic algorithm for site 1 considering all 10 GDU scenarios for both cases 1 and 2.

The heuristic algorithm resulted in one set of planting dates shown in Figure 2.9 and 10 sets of harvesting dates for 10 GDU scenarios. In other words, because each seed is harvested as soon as it accumulates its required GDU, we only need to know the planting dates of the seed populations and based on each historical GDU scenario the harvesting weeks will be known.

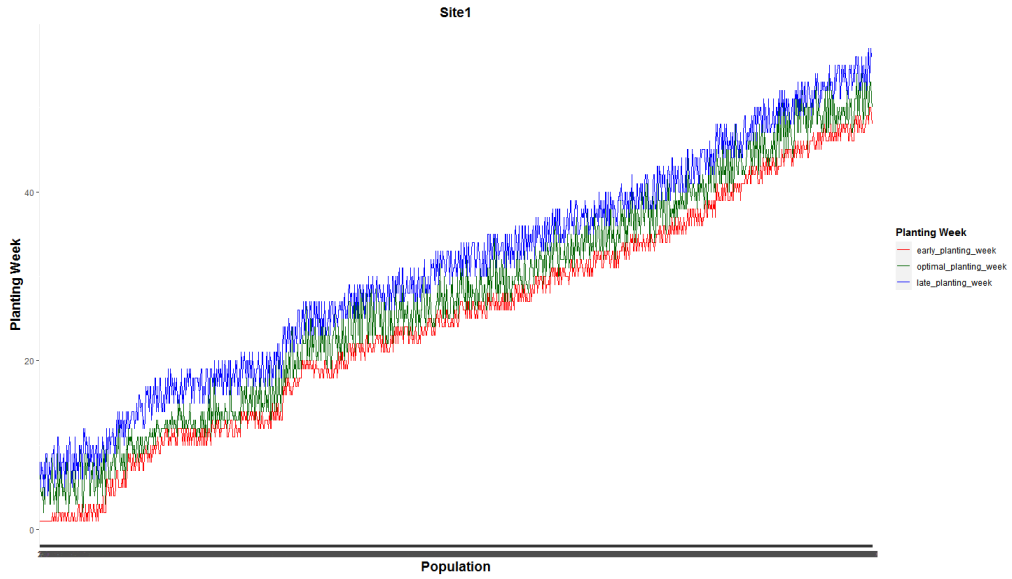


Figure 2.9: Optimal planting weeks of 1194 seed populations planted in site 1 for all 10 GDU scenarios for case 1.

Maximum weekly harvest quantities among all 10 GDU scenarios for optimal planting dates resulted from our heuristic algorithm (2.3.3.3) and for original planting dates are shown in Figure 2.10 for site 1. As it was explained in section 2.3.3.3, the heuristic algorithm finds optimal planting dates by moving planting dates of populations with minimum harvest quantities at the week with maximum harvest quantity either forward or backward. For the week with the minimum harvest quantity, the model tries to move populations from closest weeks with the larger harvest quantities to the week with minimum harvest quantity. As it is shown in Figure 2.10, weeks 17, 18, and 69 are the weeks with the minimum harvest quantities and they are causing an inconsistency in weekly harvest quantities. In Table 2.5, we manually tried to move seed populations with the minimum harvest quantities from taller bars to shorter bars to check whether the model could have improved the results. For example, the planting date of population ID 94 corresponding to week 19 was changed from week 3 to 2, and it not only did not change the weekly harvest quantities of weeks 17 and 18 but also caused inconsistency by increasing the difference between weeks 19 and 20. Moreover, it was impossible to change the planting date of population ID 1061 corresponding to

week 68 from week 52 to 53 because week 52 is the last day of its planting window. These results from Table 2.5 suggest that the planting dates resulted from our heuristic algorithm are reasonable since the changes could not improve the results.

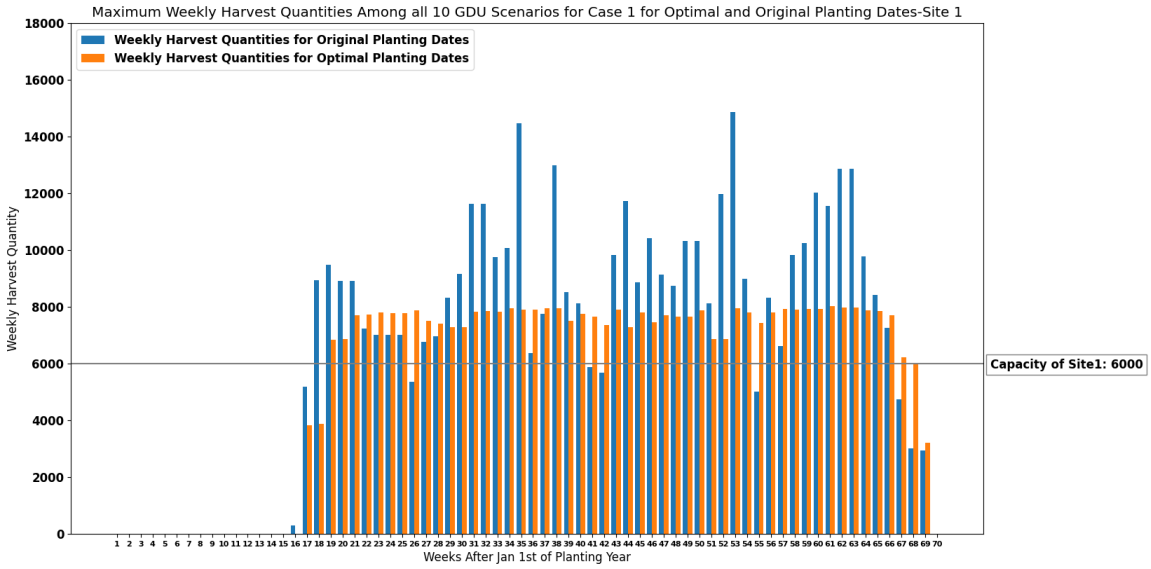


Figure 2.10: Maximum weekly harvest quantities among all 10 GDU scenarios of site 1 with a capacity of 6000 ears for case 1 for optimal and original planting dates.

The values of the sum of absolute differences between maximum weekly harvest quantities among all 10 GDU scenarios and the capacity resulted from optimal planting dates suggested by the objective function of the proposed heuristic algorithm and from original planting dates for site 1 are presented in Table 2.6. The run-time of the algorithm is also presented in Table 2.6.

2.4.3 Results of the Deterministic Model for Case 2

The same input variables as the MILP model of case 1 (2.13)-(2.16) were used in the MILP model of case 2 (2.30)-(2.33) except the number of ears produced by each population (HQ_i) planted in site 0 and site 1. As it was discussed in section 2.2 there are different quantities for the number of ears produced by each population (HQ_i) planted in site 0 and site 1 for this case in which the

model is supposed to determine the lowest capacity required for each site. Because of the complexity of this MILP model (2.30)-(2.33) we used stopping criterion of 0.1 % optimality gap for each site.

Table 2.5: Information of population IDs 94 and 1061 with the lowest harvest quantities among all populations corresponding to target weeks 19 and 68 respectively and how changing their planting dates to other weeks can affect the results.

Target Week	19	68
Population ID	94	1061
Early Planting Week	2	43
Late Planting Week	9	52
Harvest Quantity	160	68
Optimal Planting Week	3	52
New Planting Week	2	Not possible
Optimal HQs for Weeks of 17/18/19/20	3826/3881/**6845**/6869	Not applicable
HQs After Change of Planting Week for Weeks of 17/18/19/20	3826/3881/**6685**/6869	Not applicable

Table 2.6: Run-time of the proposed heuristic algorithm and the values of the sum of absolute differences between maximum weekly harvest quantities among all 10 GDU scenarios and the site capacity resulted from optimal planting dates suggested by the objective function of the proposed heuristic algorithm and from original planting dates for case 1 under the multiple GDU scenarios for site 1.

Planting date	Run-time (hours)	Sum of absolute differences
Optimal Planting	20	87421
Original Planting	-	172851

The lowest storage capacities required for each site under the predicted GDU scenario for optimal planting dates suggested by the MILP model (2.30)-(2.33) and for original planting dates are presented in Table 2.7.

Table 2.7: The lowest storage capacities required for site 0 and site 1 under the predicted GDU scenario for optimal planting dates and original planting dates.

Site	Lowest Capacity for Optimal Planting Dates	Lowest Capacity for Original Planting Dates
0	10,795	37,247
1	8,108	16,220

Figures 2.11 and 2.12 demonstrate how our proposed MILP model, represented by equations (2.30)-(2.33), for case 2 was able to determine the lowest capacity required for each site under the predicted GDU scenario while successfully scheduling planting and harvesting weeks that resulted in consistent weekly harvest quantities.

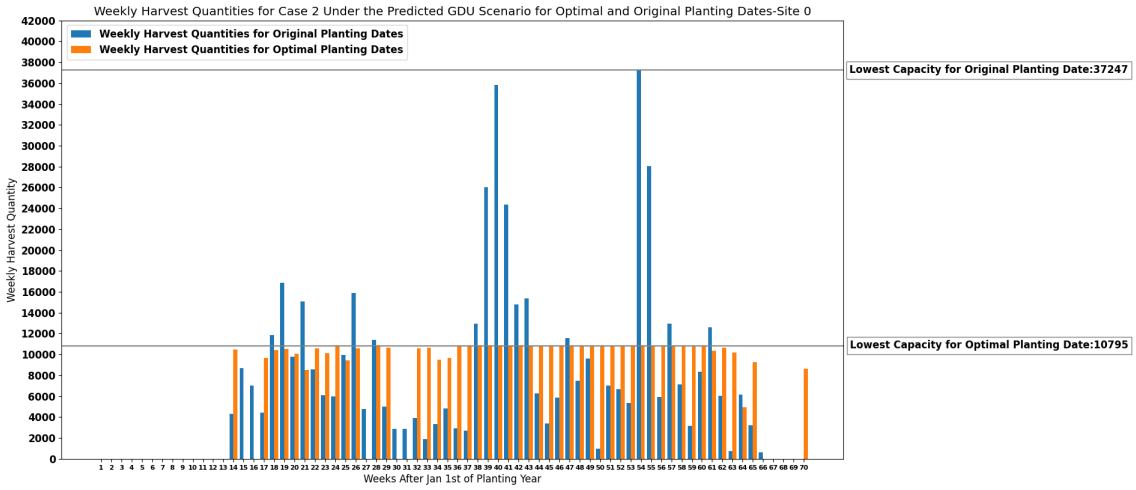


Figure 2.11: Weekly harvest quantities of site 0 with original capacity of 37247 and suggested optimal capacity of 10795 for case 2 under predicted GDU scenario.

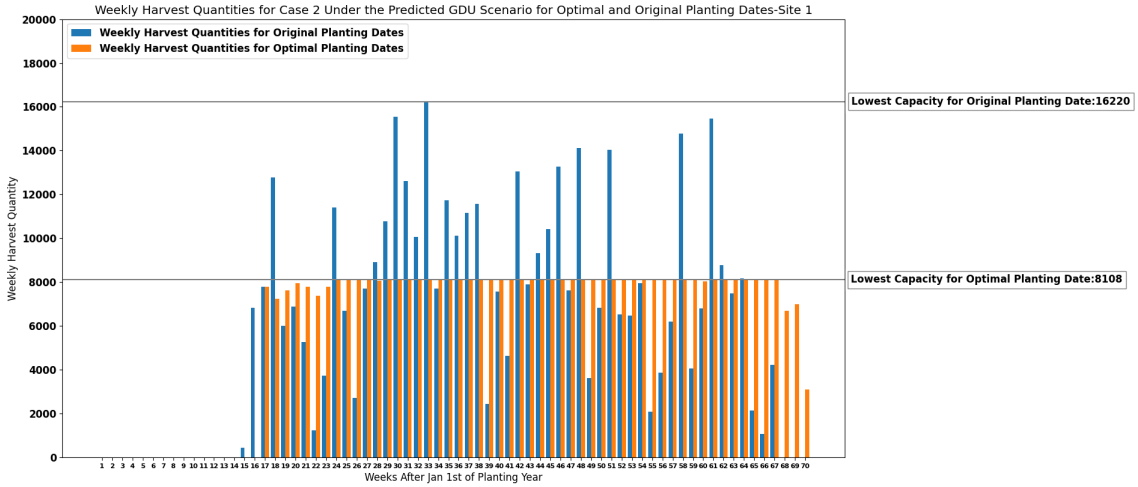


Figure 2.12: Weekly harvest quantities of site 1 with original capacity of 16220 and suggested optimal capacity of 8108 for case 2 under predicted GDU scenario.

Optimal planting weeks of 1375 seed populations planted in site 0 and 1194 seed populations planted in site 1 for case 2 under the predicted GDU scenario are shown in Figures 2.13 and 2.14 respectively. These figures also show that the optimal planting weeks suggested by our proposed MILP model (2.30)-(2.33) are between the early and late planting weeks.

2.4.4 Results of the Stochastic Model for Case 2

This subsection presents the results of our heuristic algorithm for case 2 where we need to determine the lowest required capacity for each site. As it was discussed in section 2.4.2 here we also solved case 2 only for site 1 considering multiple GDU scenarios. Figure 2.15 shows the maximum weekly harvest quantities among all 10 GDU scenarios for optimal and original planting dates and the lowest required capacities for them. In order to show that the results are reasonable we considered weeks 20, 21, and 68 and tried to move the planting dates of the populations with the lowest harvest quantities corresponding to week 20, and week 21 to one week before and the planting date of the population with the lowest harvest quantity corresponding to week 68 to one week after to see how maximum weekly harvest quantities among all 10 GDU scenarios would

change and whether these changes would give us better results. The information of those populations with the lowest harvest quantities and how changing their planting dates could affect the results are presented in Table 2.8 for target weeks of 20, 21, and 68.

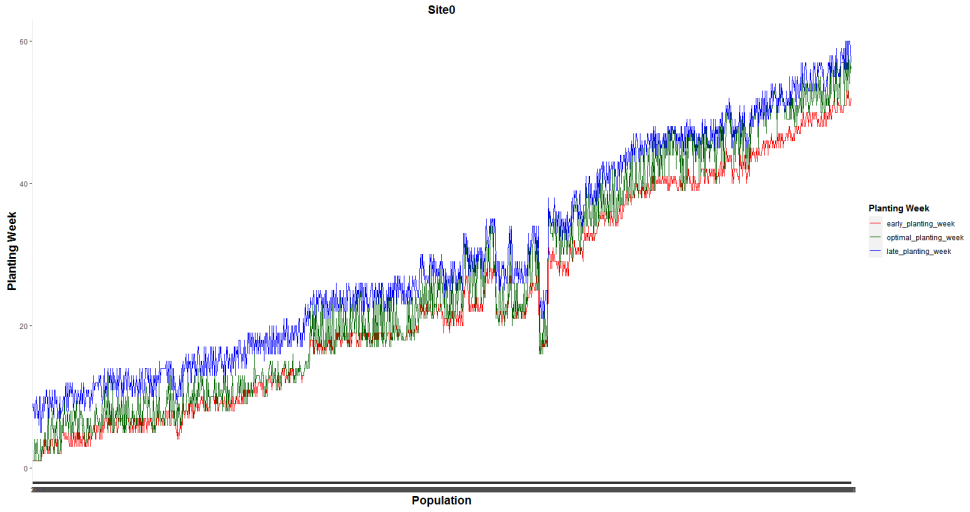


Figure 2.13: Optimal planting weeks of 1375 seed populations planted in site 0 under the predicted GDU scenario for case 2.

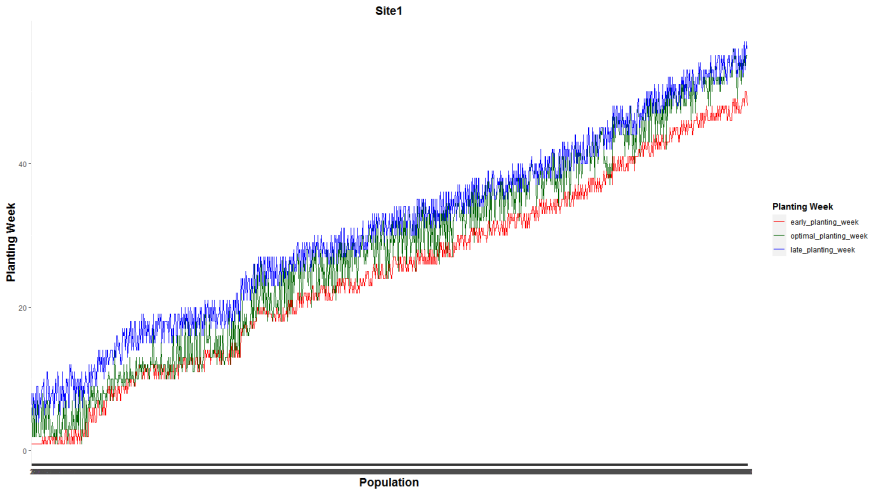


Figure 2.14: Optimal planting weeks of 1194 seed populations planted in site 1 under the predicted GDU scenario for case 2.

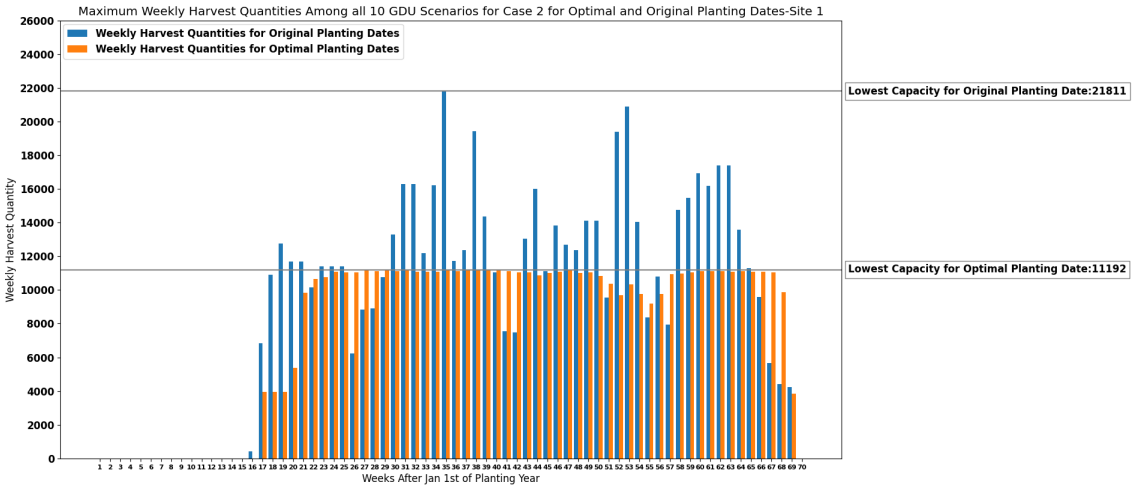


Figure 2.15: Maximum weekly harvest quantities among all 10 GDU scenarios of site 1 with original capacity of 21811 and suggested optimal capacity of 11192 for case 2.

Table 2.8: Information of the population IDs 70, 41, and 1063 with the lowest harvest quantities among all populations corresponding to target weeks 20, 21, and 68 and how changing their planting dates to other weeks can affect the results.

Target Week	20	21	68
Population ID	70	41	1063
Early Planting Week	3	1	43
Late Planting Week	7	6	52
Harvest Quantity	70	102	59
Optimal Planting Week	6	5	52
New Planting Week	5	4	Not possible
Optimal HQs for Weeks of 17/18/19/20/21, and 17/18/19/20/21/22	3958/3958/3958/5391/9851	3958/3958/3958/5391/9851/10657	Not applicable
HQs After Change of Planting Week for Weeks of 17/18/19/20/21, and 17/18/19/20/21/22	No Change	No Change	Not applicable

The lowest storage capacities required for site 1 under multiple GDU scenarios for optimal and original planting dates are presented in the following table (Table 2.9).

Table 2.9: The lowest capacities required for site 1 under multiple GDU scenarios for optimal planting dates and original planting dates.

Site	Lowest Capacity for Optimal Planting	Lowest Capacity for Original Planting
1	11192	21811

Our heuristic algorithm resulted in one set of planting dates for all 10 GDU scenarios which is shown in Figure 2.16 and 10 sets of harvesting weeks which can easily be calculated for each GDU scenario because the populations need to be harvested as soon as they earn their required GDUs.

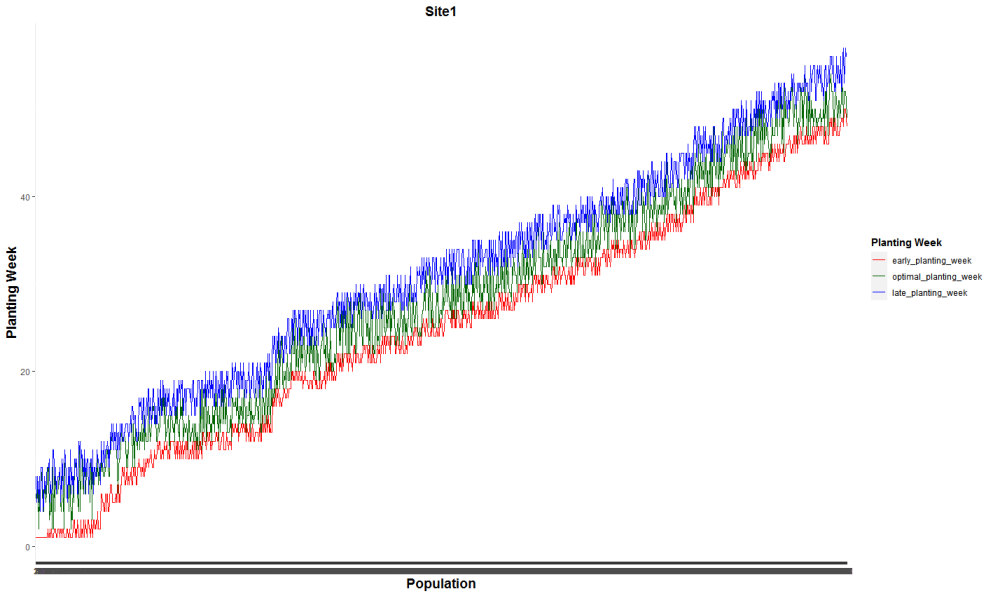


Figure 2.16: Optimal planting weeks of 1194 seed populations planted in site 1 for all 10 GDU scenarios for case 2.

2.5 Evaluation of the Proposed Heuristic Algorithm

In this section, we evaluate the performance of the proposed heuristic algorithm which was presented in section 2.3.3.3 to solve the corn scheduling problem considering multiple GDU scenarios for both storage capacity cases. To this end we solve a small problem for only storage capacity case 1 using both proposed heuristic algorithm (1) and MILP model (2.19)-(2.25). The proposed MILP model (2.19)-(2.25) with $N \times K \times T + N \times T + K \times T + T$ decision variables which is equal to 920,150 for 1194 populations (N) harvested in 70 weeks (T) under multiple GDU scenarios for 10 years (K) in site 1 can not be solved using the existing branch-and-bound algorithms due to lack of computational power. So, to evaluate the performance of the proposed heuristic algorithm (1), we created a small problem and instead of using the whole 1194 populations planted in site 1, we only used 100 populations and solved the problem using both heuristic algorithm and MILP model. The MILP model for 100 populations with 77,770 decision variables was run in MATLAB R2018a and solved with MILP commercial solver Gurobi Optimizer. The MILP solver was not able to find a feasible solution and converge even for 100 populations in 72 hours. That might be the reason the solution of the MILP model is not satisfactory compared to the solution of the heuristic algorithm proposed specifically to solve the multi GDU scenario.

Maximum weekly harvest quantities among all 10 GDU scenarios for planting dates obtained from the heuristic algorithm (1) and MILP model (2.19)-(2.25) are shown in Figure 2.17 for site 1. Run-times and objective values (sum of absolute differences between maximum weekly harvest quantities among all 10 GDU scenarios and the capacity) of the heuristic algorithm and MILP model using 100 populations from site 1 are presented in Table 2.10. As it is shown in Figure 2.17 and Table 2.10, the heuristic algorithm has better performance in terms of consistent weekly harvest quantities which are below the capacity and run time.

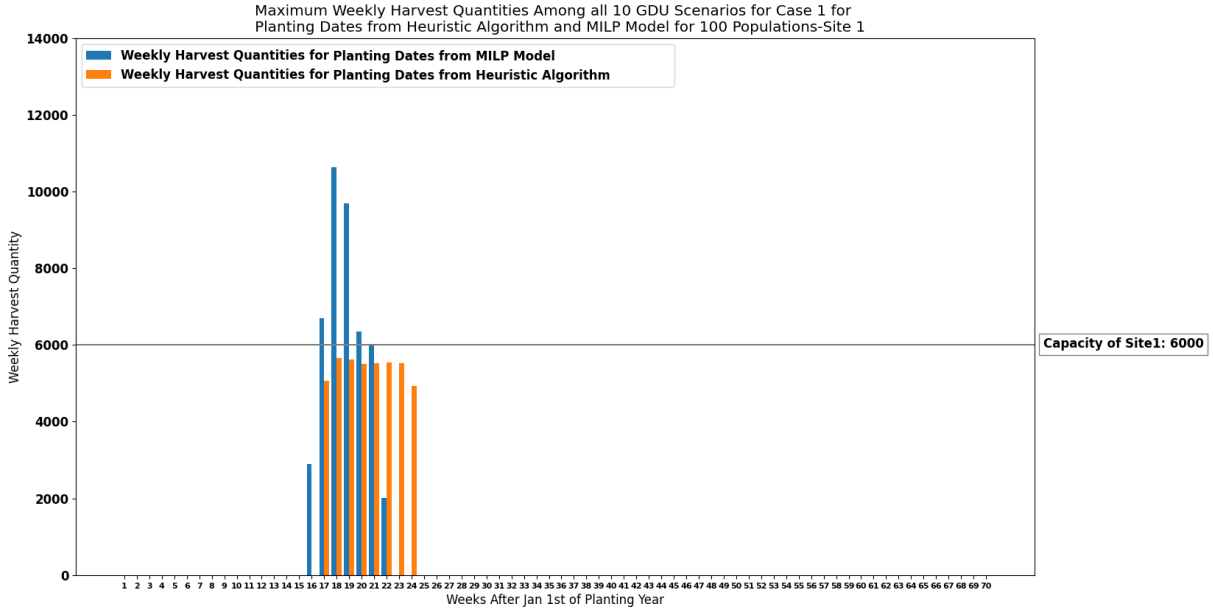


Figure 2.17: Maximum weekly harvest quantities among all 10 GDU scenarios of site 1 with a capacity of 6000 ears for case 1 for planting dates obtained from the heuristic algorithm and MILP model considering 100 populations. The solution from the MILP model violated the capacity constraint because the MILP solver was unable to find a feasible solution within the 72-hour time limit.

Table 2.10: Run-times and objective values (sum of absolute differences between maximum weekly harvest quantities among all 10 GDU scenarios and the capacity) of the heuristic algorithm and MILP model using 100 populations from site 1.

Planting date	Run-time (hours)	Sum of absolute differences
Heuristic Algorithm	0.6	4614
MILP Model	72	16491

2.6 Discussion and Conclusion

In this paper, we proposed two MILP models and a heuristics algorithm to help growers and farmers schedule planting and harvesting dates of different corn populations to have consistent harvest quantities that are below the storage capacity of the site for two storage capacity cases considering a deterministic GDU scenario and multiple GDU scenarios together.

Considering a deterministic GDU scenario, the results of the proposed MILP models show that the models can successfully schedule the planting and harvesting dates of diverse seed populations with various planting windows, required GDUs, and harvest quantities planted in two sites in a way that there is no overflow of the capacity for case 1 and a consistent weekly number of ears for both cases. As it was shown in Figures 2.5, 2.6, 2.11, and 2.12 the models also could improve the results over the original planting dates provided by the challenge in terms of avoiding overflowing capacity for case 1, suggesting much less lowest required capacity for case 2, and consistency of weekly harvest quantities for both cases. As it was mentioned before having no carry over the storage capacity can help farmers avoid having to dump harvested crops or leave crops unharvested. The results assure that the proposed MILP models can maximize the benefit by having no carry over of the storage capacities for case 1 and decreasing the lowest required capacities from 37247 and 16220 using original planting dates to 10795 and 8108 using optimal planting dates for site 0 and site 1 respectively for case 2.

Running our MILP models considering a deterministic GDU scenario (predicted GDU) for different amounts of GDUs indicates that the low average temperature or low daily GDUs makes the optimization models infeasible and prevents us from harvesting the whole populations in 70 weeks as the corn populations can not accumulate their required GDU to reach full maturity. Moreover, the results of our proposed MILP models indicate that different weather conditions or GDU quantities affect the number of harvesting weeks and harvest quantities. These explain why we proposed a heuristic algorithm to solve the problem considering multiple GDU scenarios at the same time. Additionally, the results from the proposed MILP model for case 2 considering a deterministic GDU scenario reveal that the amount of GDU units or weather condition also affects

the lowest capacity required and the lower GDU units (lower average temperature) resulted in higher capacity required. It is due to the fact that corn populations accumulate their required GDUs slower, and as a result, higher number of corn populations should be harvested in later weeks. Therefore, higher storage capacity is required due to the limited number of harvesting weeks. As it was described in section 2.3.3 the proposed MILP models considering multiple GDU scenarios for both cases have $N \times K \times T + N \times T + K \times T + T$ decision variables which is equal to 1,059,520 and 920,150 variables for site 0 and site 1 respectively and they require considerable amount of computational power which makes it infeasible using exact algorithms (N, K, and T are the number of seed populations in each site, number of years of GDU scenarios, and number of harvesting weeks respectively). As such, we proposed a new heuristic algorithm based on simulated annealing to solve the problem. The proposed heuristic algorithm only has N decision variables which is equal to the number of seed populations in each site (1375 for site 0 and 1194 for site 1) and took 20 hours to solve each case in Python. To evaluate the performance of the proposed heuristic algorithm, we solved a small problem with 100 seed populations for only storage capacity case 1 using both proposed heuristic algorithm (1) and MILP model (2.19)-(2.25). The results show that even for a small problem the proposed heuristic algorithm not only has lower computational time but also it has better performance in terms of consistent weekly harvest quantities which are below the capacity.

The different results from cases 1 and 2 raise an interesting question of cost-benefit analysis. On the one hand, case 2 requires storage capacities of 10795 (site 0) and 8108 (site 1) considering predicted GDU scenario and 11192 (site 1) considering multiple GDU scenarios, compared with the capacities of 7000 (site 0) and 6000 (site 1) for case 1. On the other hand, the higher capacity in case 2 allows for a total harvest quantity of 923412 (sites 0 and 1 combined), compared with the total harvest quantity of 657546 in case 1. Given the cost of a higher storage capacity and the benefit of an increase harvest quantity, we would be able to determine if case 2 is economically more beneficial than case 1.

Since the proposed planting date scheduling models depend on the GDU prediction model performance, we suggest to improve the accuracy of the GDU prediction model with advanced machine learning models such as transformers. Additionally, as it was mentioned in previous sections one of the assumptions of the 2021 Syngenta crop challenge was to harvest the corn hybrids as soon as they reach their required GDUs. For the future work the models can be modified to allow corns to be harvested up to a certain time after they reach maturity rather than being immediately harvested. Moreover, another important criterion which was not required for the challenge and should be taken into account is having consecutive harvesting weeks and the model can be modified to produce the harvesting weeks that are consecutive. We hope that our work leads to the advancement of crop planting and harvest scheduling to benefit plant science as a whole.

Acknowledgements

This work was supported by USDA under NIFA program (grant number 2017-67007-26175 /accession number 1011702) and NSF under LEAP HI and GOALI programs (grant number 1830478) and EAGER program (grant number 1842097). This work was also supported by the Plant Sciences Institute at Iowa State University. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the funding agencies. We also thank Syngenta for providing the valuable datasets.

Data availability

The data analyzed in this study was provided by Syngenta AG company for 2021 Syngenta Crop Challenge. We accessed the data through annual Syngenta Crop Challenge. During the challenge, September 2020 to January 2021, the data was open to the public. Data cannot be shared publicly because of non-disclosure agreement. Data are available from Syngenta (contact via https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php) for researchers who meet the criteria for access to confidential data.

2.7 References

Ahuja, R. (1985). Minimax linear programming problem. *Operations Research Letters*, 4(3):131–134.

Bachmann, J. (2008). *Scheduling vegetable plantings for continuous harvest*. ATTRA.

Bertsimas, D., Tsitsiklis, J., et al. (1993). Simulated annealing. *Statistical science*, 8(1):10–15.

Cid-Garcia, N. M., Bravo-Lozano, A. G., and Rios-Solis, Y. A. (2014). A crop planning and real-time irrigation method based on site-specific management zones and linear programming. *Computers and electronics in agriculture*, 107:20–28.

Ferguson, T. S. (2000). Linear programming: A concise introduction. http://web.tecnico.ulisboa.pt/mcasquilho/acad/or/ftp/FergusonUCLA_LP.pdf.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Khaki, S., Khalilzadeh, Z., and Wang, L. (2019). Classification of crop tolerance to heat and drought—a deep convolutional neural networks approach. *Agronomy*, 9(12):833.

Khaki, S., Khalilzadeh, Z., and Wang, L. (2020a). Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach. *Plos one*, 15(5):e0233382.

Khaki, S., Pham, H., Han, Y., Kent, W., and Wang, L. (2020b). High-throughput image-based plant stand count estimation using convolutional neural networks. *arXiv preprint arXiv:2010.12552*.

Khaki, S., Pham, H., Han, Y., Kuhl, A., Kent, W., and Wang, L. (2020c). Convolutional neural networks for image-based corn kernel detection and counting. *Sensors*, 20(9):2721.

Khaki, S., Pham, H., Han, Y., Kuhl, A., Kent, W., and Wang, L. (2020d). Deepcorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation. *arXiv preprint arXiv:2007.10521*.

Khaki, S., Pham, H., and Wang, L. (2020e). Yieldnet: A convolutional neural network for simultaneous corn and soybean yield prediction based on remote sensing data. *arXiv preprint arXiv:2012.03129*.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621.

Khaki, S., Wang, L., and Archontoulis, S. V. (2020f). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750.

Lawler, E. L. and Wood, D. E. (1966). Branch-and-bound methods: A survey. *Operations research*, 14(4):699–719.

Meyers, K. and Rhode, P. (2020). Yield performance of corn under heat stress: A comparison of hybrid and open-pollinated seeds during a period of technological transformation, 1933-1955. Technical report, National Bureau of Economic Research.

National Corn Handbook (2020). https://www.extension.purdue.edu/extmedia/nch/nch-40.html.

Sarker, R. A., Talukdar, S., and Haque, A. A. (1997). Determination of optimum crop mix for crop cultivation in bangladesh. *Applied Mathematical Modelling*, 21(10):621–632.

Syngenta Crop Challenge (2021). https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php.

The North Dakota Agricultural Weather Network (NDAWN) (2020). https://ndawn.ndsu.nodak.edu/help-corn-growing-degree-days.html.

Van Laarhoven, P. J. and Aarts, E. H. (1987). Simulated annealing. pages 7–15.

Wright, H. (1980). Commercial hybrid seed production. *Hybridization of crop plants*, pages 161–176.

CHAPTER 3. A HYBRID DEEP LEARNING-BASED APPROACH FOR OPTIMAL GENOTYPE BY ENVIRONMENT SELECTION

Zahra Khalilzadeh[1], Motahareh Kashanian[2], Saeed Khaki[3], and Lizhi Wang[4]

[1]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[2]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[3]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[4]School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74078, USA

Submitted to Frontiers in Artificial Intelligence, section AI in Food, Agriculture and Water.

Abstract

Accurately predicting crop yield is vital for enhancing agricultural production and its resiliency in diverse climatic conditions. Integrating weather data throughout the crop growing season, especially for various genotypes, is crucial for these predictions. It represents a significant stride in comprehending how climate change affects a variety's adaptability. In the MLCAS2021 Crop Yield Prediction Challenge, the Third International Workshop on Machine Learning for Cyber-Agricultural Systems released a dataset for soybean hybrids consisting of 93,028 training performance records to predict yield for the 10,337 testing performance records. This dataset spans 159 locations across 28 states in the U.S. and Canadian provinces over a 13-year period, from 2003 to 2015. It comprises details on 5,838 distinct genotypes and daily weather data for a 214-day growing season, encompassing all possible location and year combinations. As one of the winning teams, we design two novel convolutional neural network (CNN) architectures. The first proposed model combines CNN and fully-connected (FC) neural networks (CNN-DNN model). The second proposed model adds a long short-term memory

(LSTM) layer at the end of the CNN part for the weather variables (CNN-LSTM-DNN model). We utilize the Generalized Ensemble Method (GEM) to determine the optimal weights of the proposed CNN-based models to achieve higher accuracy than other baseline models. The GEM model we introduce demonstrates superior performance compared to all other baseline models employed in soybean yield prediction. It exhibits a lower Root Mean Square Error (RMSE) ranging from 5.55% to 39.88%, a reduced Mean Absolute Error (MAE) ranging from 5.34% to 43.76%, and an improved coefficient ranging from 1.1% to 10.79% in comparison to the baseline models when evaluated on test data. The proposed CNN-DNN model is then employed to identify the best-performing genotypes for various locations and weather conditions, making yield predictions for all potential genotypes in each specific setting. The dataset provides unique genotype information on seeds, allowing investigation of the potential of planting genotypes based on weather variables. The proposed data-driven approach can be valuable for genotype selection in scenarios with limited testing years. We also perform a feature importance analysis utilizing RMSE change to identify crucial predictors impacting our model's predictions. The location variable exhibits the highest RMSE change, emphasizing its pivotal role in predictions, followed by maturity group (MG), year, and genotype, showcasing their significance during crop growth stages and across different years. In the weather category, maximum direct normal irradiance (MDNI) and average precipitation (AP) display higher RMSE changes, indicating their importance. In addition, we explore the impact of incorporating state-level soil data alongside the variables from the MLCAS2021 Crop Yield Prediction Challenge. However, the current data constraints pose a limitation, as the available data includes only location IDs and states without latitude and longitude details. This constraint prevents us from having specific soil variables for each location ID, necessitating the use of uniform soil variables for all locations within the same state. Consequently, the lack of exact geographical coordinates for each location ID restricts our spatial information to state-level knowledge. Despite these constraints, our findings suggest that the integration of soil variables does not substantially enhance the predictive capabilities of the models under the present data conditions.

3.1 Introduction

The world's population is projected to reach almost 10 billion by 2050 [Nations et al., 2017], and climate change is expected to have a significant impact on crop yields in the coming years. As a result, there is an urgent need to increase crop production in order to feed the growing population. The current global food production systems are facing several challenges such as the increasing frequency and severity of droughts, floods, heatwaves and increased pests and diseases, which are all associated with climate change [Kumar, 2016]. These challenges are likely to affect crop yields and food security, making it essential to develop new strategies to increase crop production.

One of the main strategies for increasing crop production is to develop climate-resilient crops through breeding programs. This involves selecting and crossbreeding plants that are better able to withstand the effects of climate change, such as drought or heat stress. Despite the focus on climate resilience in breeding programs, there is mounting evidence of the difficulties and challenges in creating crops capable of handling the effects of climate change. These challenges stem from the contradiction between the pressing need for breeding in response to climate change and the inadequate understanding of how genotype and environment interact with each other [Xiong et al., 2022]. Another approach is to use crop simulation models that integrate environmental information and tools into the breeding analysis process to tackle the effects of climate change and anticipate crop growth and yield under different climate scenarios [de Los Campos et al., 2020, Heslot et al., 2014]. However, crop simulation models have limitations such as complexity, where the simulations may not be able to fully capture all the interactions of multiple factors such as genetics, environment and management practices, leading to inaccurate predictions. Additionally, data availability, validation, computational resources, the limitation of analyzing a limited number of genotypes, and simplification of reality in the models are other limitations of simulation crop modeling [Roberts et al., 2017, Hajjarpoor et al., 2022]. To overcome the limitations of crop growth models, studies are emerging recently to utilize statistical methods as promising alternatives and complementary tools. Among these methods, Machine Learning (ML) is a practical statistical approach that has gained popularity due to advancements in big-data technologies and

high-performance computing. ML algorithms can help farmers to increase crop production in response to climate change by providing capabilities such as crop yield prediction [Shahhosseini et al., 2021, Khaki and Wang, 2019], climate change impact modeling Crane-Droesch [2018], climate-smart crop breeding Xu et al. [2022], automation of farming equipment Patil and Thorat [2016], market price prediction Chen et al. [2021], water management optimization Lowe et al. [2022], disease and pest forecasting [Domingues et al., 2022], and precision agriculture Sharma et al. [2020]. These capabilities can help farmers to plan for and adapt to changing weather patterns, identify resilient crops, optimize crop management practices, and make better decisions to increase crop production. The challenge of effectively training ML algorithms is posed by the inconsistent spatial and temporal data regarding some of the production and management inputs, such as planting date, fertilizer application rate, and crop-specific data. This is a problem that needs to be addressed for efficient ML algorithm training.

Genotype by environment interaction is a challenging factor that limits the genotype selection for increased crop yields in unseen and new environments especially with the presence of global climate change. Plant breeders typically choose hybrids based on their desired traits and characteristics, such as yield, disease resistance, and quality. They first select parent plants with desirable traits and cross them to create a new hybrid. The new hybrids are then tested in various environments to determine their performance, finally the hybrids with the highest yield are selected [Bertan et al., 2007]. However, this approach can be extremely time-consuming and tedious due to the vast number of possible parent combinations that require testing [Khaki et al., 2020a]. This highlights the importance of having a data driven approach to select genotypes with the highest performance in response to climates as well as other environmental variables using limited years of field testing per genotype. For example, Arzanipour and Olafsson [2022], suggests employing imputation methods to address the issue of incomplete data, particularly when certain crop types are not cultivated in every observed environment. This perspective views these absent data points not merely as traditional missing values but as potential opportunities for additional observations. In this study, we introduce a new deep learning framework for predicting crop yields using

environmental data and genotype information. The framework is designed to identify the most efficient genotype for each location and environment, by first forecasting crop yields based on the given weather conditions in each location for all available genotypes, and then selecting the optimal genotype with the highest yield in each specific location and environmental scenario. This strategy helps in enhancing policy and agricultural decision-making, optimizing production, and guaranteeing food security. To the best of our knowledge this is the first study to use a deep learning approach for optimal genotype by environment selection.

Over the years, several machine learning algorithms have been employed for predicting performance of crops under different environmental conditions. These include Convolutional Neural Network (CNN) [Srivastava et al., 2022], Long Short Term Memory (LSTM) networks [Shook et al., 2021], Regression Tree (RT) [Veenadhari et al., 2011], Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Extreme Gradient Boosting (XGBoost), Least Absolute Shrinkage and Selection Operator (LASSO) [Kang et al., 2020], and Deep Neural Network (DNN) [Khaki and Wang, 2019]. In time series prediction tasks, deep neural networks have proven to be robust to inputs with noise and possess the ability to model complex non-linear functions Dorffner [1996]. By utilizing deep learning models, it becomes possible to tackle complex data, as these models can effectively learn the non-linear relationships between the multivariate input data, which includes weather variables, maturity group/cluster information, genotype information, and the predicted yield.

Our proposed hybrid CNN-LSTM model consists of CNNs and LSTM. CNNs can handle data in multiple array formats, such as one-dimensional data like signals and sequences, two-dimensional data such as images, and three-dimensional data like videos. A typical CNN model consists of a series of convolutional and pooling layers, followed by a few fully connected (FC) layers. There are several design parameters that can be adjusted in CNNs, including the number of filters, filter size, type of padding, and stride. Filters are weight matrices used to process the input data during convolution. Padding involves adding zeroes to the input data to maintain its dimensional structure, while the stride refers to the distance by which the filter is moved during processing

[Albawi et al., 2017]. Recurrent Neural Networks (RNNs) are a type of deep learning model designed for handling sequential data. The key advantage of RNNs is their ability to capture time dependencies in sequential data due to their memory mechanism, allowing them to use information from previous time steps in future predictions [Sherstinsky, 2020, Lipton et al., 2015]. LSTM networks are a specialized type of RNNs that address the issue of vanishing gradients in traditional RNNs [Hochreiter and Schmidhuber, 1997, Sherstinsky, 2020]. LSTMs are particularly beneficial for capturing long-term dependencies in sequential data, and they maintain information for longer periods of time compared to traditional RNNs [Hochreiter and Schmidhuber, 1996]. These characteristics make LSTMs highly effective for handling data with complex temporal structures, such as speech and video [Xie et al., 2019, Li et al., 2019]. Furthermore, LSTMs have been successfully utilized in multivariate time series prediction problems [Shook et al., 2021, Sun et al., 2019, Gangopadhyay et al., 2018], and they are flexible and handle varying length inputs, making them suitable for processing sequential data with different lengths [Sutskever et al., 2014].

Crop yield prediction has been more recently improved by the application of deep learning methods. Khaki and Wang [2019] utilized deep neural networks to predict corn yield for various maize hybrids using environmental data and genotype information. Their study involved designing a deep neural network model that could forecast corn yield across 2,247 locations from 2008 to 2016. With regards to the accuracy of their predictions, the model they developed outperformed others such as LASSO, shallow neural networks, and regression trees, exhibiting a Root Mean Square Error (RMSE) of 12% of the average yield when using weather data that had been predicted, and an RMSE of 11% of the average yield when using perfect weather data. Environmental data including weather and soil information and management practices were used as inputs to the CNN-RNN model developed by Khaki et al. [2020b] for corn and soybean yield prediction across the entire Corn Belt in the U.S. for the years 2016, 2017, and 2018. Their proposed CNN-RNN model outperformed other models tested including RF, deep fully connected neural networks, and LASSO, achieving a notable improvement with an RMSE of 9% and 8% for corn and soybean average yields, respectively. They also employed a guided backpropagation technique to select features and

enhance the model's interpretability. Similarly, Sun et al. [2019] adopted a comparable strategy, utilizing a CNN-LSTM model to predict county-level soybean yields in the U.S. using satellite imagery, climate data, and other socioeconomic factors. Their results show that the CNN-LSTM model can capture the spatiotemporal dynamics of soybean growth and outperform other models in terms of accuracy and computational efficiency. Oikonomidis et al. [2022] utilized a publicly available soybean dataset, incorporating weather and soil parameters to develop several hybrid deep learning-based models for crop yield prediction. Comparing their models with the XGBoost algorithm, the authors found that their hybrid CNN-DNN model outperformed the other models with an impressive RMSE of 0.266, Mean Squared Error (MSE) of 0.071, and Mean Absolute Error (MAE) of 0.199. However, none of these studies have addressed the issue of determining which crop genotype to plant based on the given weather conditions. The dataset, which was developed, prepared, and cleaned by Shook et al. [2021], provided us with unique genotype information on seeds, allowing us to investigate the potential of planting genotypes based on weather variables. Our proposed data-driven approach can be particularly valuable for selecting optimal genotypes when there are limited years of testing available. This is because the traditional approach of selecting the best genotypes based on a small number of years of field trials can be unreliable due to variations in weather and other environmental factors. By leveraging large datasets with genotype and weather information, it becomes possible to develop more accurate models that can predict the performance of different genotypes in various weather conditions. This can ultimately lead to the identification of genotypes that are both high-yielding and adaptable to different environments. Given that land for agriculture is limited, such data-driven approaches can help improve the productivity of crops per acre, as well as the quality and productivity of food crops through plant breeding.

This study has four main objectives. Firstly, it proposes two novel CNN architectures that incorporate a 1-D convolution operation and an LSTM layer. To achieve higher accuracy than other baseline models, the Generalized Ensemble Method (GEM) is utilized to determine the optimal weights of the proposed CNN-based models. Secondly, the proposed CNN-DNN model is utilized to select optimal genotypes for each location and weather condition. This is achieved by

predicting the yield for all possible genotypes in each specific location and environmental scenario. Thirdly, the study assesses the impact of location, maturity group (MG), genotype, and weather variables on prediction outcomes, investigating critical time periods for weather variables in yield predictions throughout the growing season of 30 weeks. Lastly, the study investigates the impact of soil variables on Soybean yield prediction by incorporating state-level soil variables. Through these objectives, this study demonstrates the value of using data-driven approaches in plant breeding and crop productivity research. Figure 3.1 presents a visual representation of the paper's objectives and outlines the conceptual framework adopted in this study.

The structure of this paper is as follows. Section 3.2 introduces the dataset used in this study. In Section 3.3, we propose a methodology for crop yield prediction and optimal genotype selection using two CNN-based architectures with a 1-D convolution operation and LSTM layer, as well as the GEM to find optimal model weights. This section also includes implementation details of the models used in this research, along with the design of experiments. Section 3.4 presents the experimental results, followed by an analysis of the findings in Section 3.5. Finally, in Section 3.6, we conclude the paper by discussing the contributions of this work and highlighting potential avenues for future research.

3.2 Data

In this paper, the data analyzed was taken from the MLCAS2021 Crop Yield Prediction Challenge (MLCAS, 2021) and consisted of 93,028 training and 10,337 testing performance records from 159 locations across 28 states in the U.S. and Canadian provinces, over 13 years (2003 to 2015). The data included information on 5,838 unique genotypes and daily weather data for a 214-day growing season, covering all location and year combinations. This data was prepared and cleaned by Shook et al. [2021]. The unique characteristic of this dataset is that it enables us to capture the biological interactions complexity, and temporal correlations of weather variables, as it provides both daily weather variables during the growing season for different locations and genotype data. The dataset included a set of variables for each performance record, which are as follows:

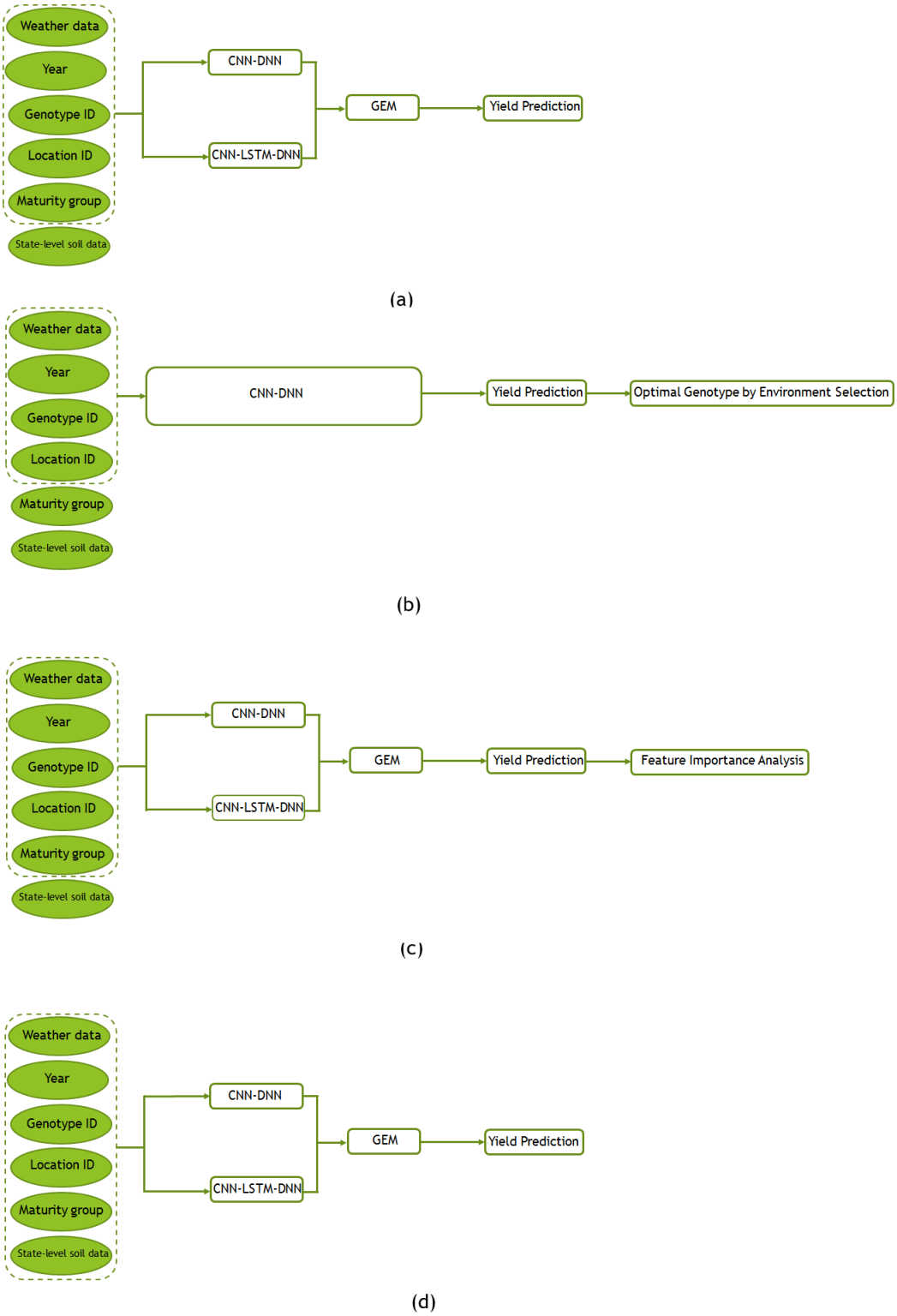


Figure 3.1: Conceptual framework of this study's objectives. Each component (a), (b), (c), and (d) corresponds to specific subsections in the paper: Subsection 3.3.2, Subsection 3.3.3, Subsection 3.5.1, and Subsection 3.5.2, respectively.

• Weather: Every performance record in the dataset included a multivariate time-series data for 214 days, which represent the crop growing season between April 1st and October 31st. Each day in the record contained seven weather variables, including average direct normal irradiance (ADNI, Wm^{-2}), average precipitation (AP, inches), average relative humidity (ARH, Percentage), maximum direct normal irradiance (MDNI, Wm^{-2}), maximum surface temperature (MaxSur, °C), minimum surface temperature (MinSur, °C), and average surface temperature (AvgSur, °C). Records with the same location and yield year share the same set of weather variables.

• Maturity group: The dataset included 10 maturity groups corresponding to different regions.

• Genotype IDs: The dataset contained 5,838 distinct genotypes, which were further clustered into 20 groups using the K-means clustering technique as described in Shook et al. [2021]. The resulting hard clustering approach allowed us to obtain a unique cluster ID for each of the 5839 genotypes in the dataset.

• State: The state information was provided for each performance record, indicating the specific state that the record corresponds to. The data covers 28 U.S. states and Canadian provinces in total.

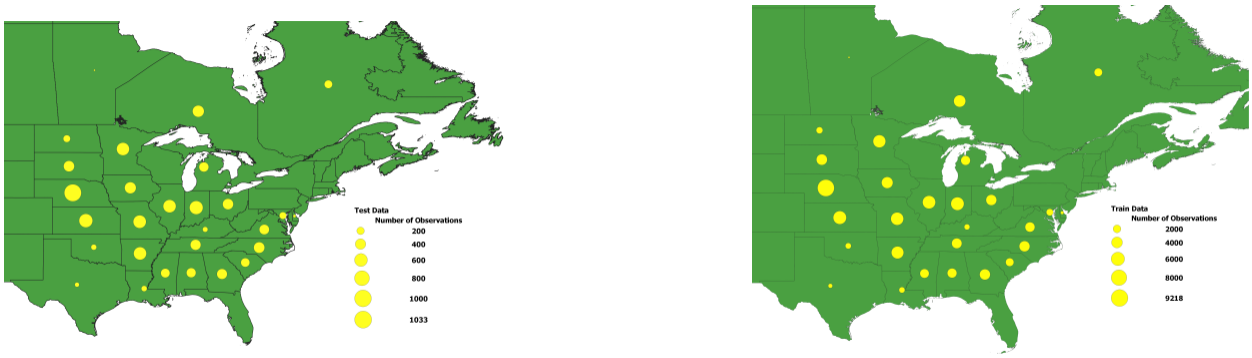
• Location ID: For each performance record, the dataset included the corresponding location ID, indicating the unique identifier for the location associated with the record. The data was collected from a total of 159 locations.

• Year: The performance record dataset contained information on the year when the yield was recorded, ranging from 2003 to 2015.

• Yield: The yield performance dataset included the observed average yield of soybean in bushels per acre across the locations in 28 U.S. states and Canadian provinces, between the years 2003 and 2015.

The goal of the 2021 MLCAS Crop Yield Prediction Challenge was to predict soybean yield for the test data consisting of 10,337 performance records including observations from all years and locations. As the competition did not provide the ground truth response variables for the test data, our analysis in this paper relied solely on the training dataset, comprising 93,028 samples. Figure

3.2 displays the distribution of performance records across 28 U.S. states and Canadian provinces in the test and train datasets. The size of each yellow dot corresponds to the size of the dataset for the corresponding state\province. Table 3.1 provides an overview of the summary statistics for both the dependent variable, soybean yield, and all independent variables used in the study(only training dataset).



(a) (b)

Figure 3.2: The distribution of performance records across 28 U.S. states and Canadian provinces in the test (a) and train (b) datasets. The size of each yellow dot corresponds to the size of the dataset for the corresponding state\province.

3.3 Method

3.3.1 Data Preprocessing

The pre-processing tasks were conducted to ensure the data is in a useful and efficient format for fitting machine learning models. One of the main tasks involved one-hot encoding the categorical variables, which included maturity group, year, location IDs, and genotype IDs. For the genotype data we tried both genotype clusters and the unique genotypes. The results demonstrated a significant improvement when the genotype IDs were included with other variables. In one-hot encoding, each unique value of each categorical variable is represented as a new binary feature in a

Table 3.1: Summary statistics of soybean yield data. The unit of yield is bushels per acre.

Summary Statistics	**Value**
Total number of locations	159
Year range	2003-2015
Mean yield	50.66
Standard deviation of yield	15.95
25th percentile of yield	39.8
Median yield	50.60
75th percentile of yield	61.40
Minimum yield	0.4
Maximum yield	112.40
Number of weather components	7
Number of maturity groups	10
Number of genotype IDs	5,838
Number of observations	93,028

new column. This means that for every observation, a value of 1 is assigned to the feature that corresponds to its original category, while all other features are set to 0. This technique results in a new binary feature being created for each possible category, allowing for more accurate modeling and prediction.

To reduce the complexity of the daily weather data and make it more suitable for analysis, we aggregated the feature values by taking the average and downsampling the data to a 4-day level. As a result of this downsampling and feature aggregation, we were able to reduce the number of model parameters significantly, with a dimension reduction ratio of 214:53. Reducing the daily weather data to a weekly level through downsampling has been commonly utilized in yield prediction studies to address the issue of excessive granularity in the data. This practice has been validated in prior research studies [Khaki and Wang, 2019, Shook et al., 2021, Srivastava et al., 2022].

Given the diverse range of values and varying scales of weather variables, it is important to avoid bias that may arise from a single feature. To address this, we applied the z-score normalization technique (Equation 3.1) to standardize all weather variable values. This technique rescales all weather variables to conform to a standard normal distribution, preventing any

unintended bias on the results. In addition to mitigating bias, standardizing the weather variable values also improves the numerical robustness of the models and accelerates the training speed.

$$W_{i,j} = \frac{w_{i,j} - \bar{w}_j}{\sigma_j} \tag{3.1}$$

where $W_{i,j}$ is the standardized value of the ith observation of the jth weather variable (j ranges from 1 to K, where K represents the total number of weather variables, which in this case is 371 (7 variables * 53 time periods)), $w_{i,j}$ is the original value of the ith observation of the jth weather variable, $\bar{w}_j$ is the mean of the jth weather variable, and σ_j is the standard deviation of the jth weather variable. The formula rescales each variable to have a mean of 0 and a standard deviation of 1.

3.3.2 Model development

In this section, we introduce two proposed models, CNN-DNN and CNN-LSTM-DNN, for predicting crop yield using location, MG, genotype, and weather data. These models are designed to handle the temporal features of weather data, which play a crucial role in crop yield prediction. CNN-DNN is a combination of CNNs and Deep Neural Networks (DNNs), while CNN-LSTM-DNN is a combination of CNNs, LSTM networks, and DNNs. Both models are trained and evaluated using the same dataset.

To improve the accuracy of our yield predictions, we propose using the GEM method that combines the predictions of both models. This approach allows us to leverage the strengths of each model and obtain better RMSE values than either model alone. In the following subsections, we describe the architecture and training procedures for the CNN-DNN and CNN-LSTM-DNN models, as well as the implementation of the GEM method for the yield prediction.

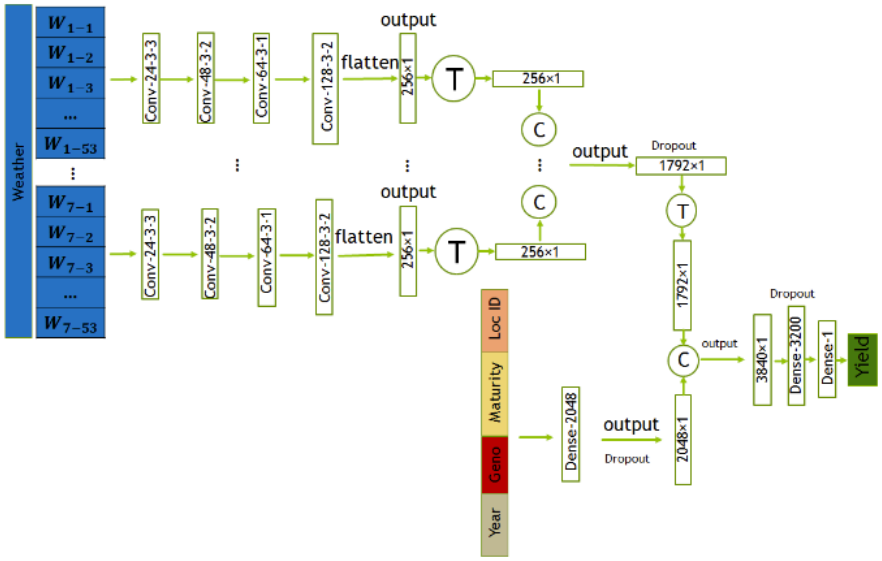
3.3.2.1 Proposed CNN-DNN Model

The first proposed model architecture combines CNNs and fully−connected (FC) neural networks. The weather variables measured throughout the growing season are taken as input in the convolutional neural network part of the model, which captures their temporal dependencies, and

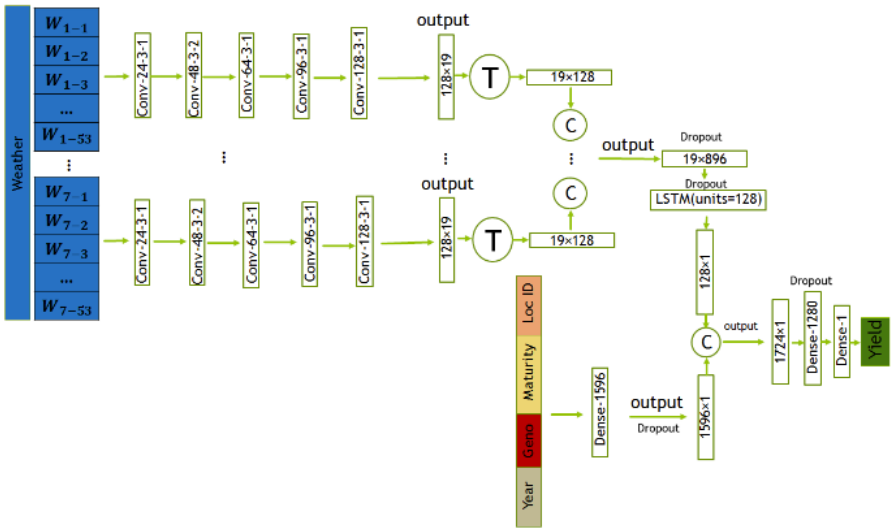
linear and nonlinear effects through 1−dimensional convolution operations. The CNN part of the model takes in the seven weather variables separately and concatenates their corresponding output for capturing their high−level features. The data for genotype, maturity group, location, and year (input _others) are fed into a fully−connected neural network with one layer. The high−level features from the CNN are then combined with the output of the fully−connected neural network for input _others data. The combined features are then processed through two additional FC layers before yielding the final prediction of the soybean yield. Moreover, to prevent overfitting, three dropout layers with dropout ratios of 0.5, 0.7, and 0.2 are respectively added to the fully connected layer after the CNN layer, at the end of the fully connected layer for input _others data, and at the final layer of the model. The proposed modeling architecture is designed to capture the complex interactions between weather data, genotype IDs, maturity groups, year, and location IDs for an accurate yield prediction and is illustrated in Figure 3.3.

3.3.2.2 Proposed CNN-LSTM-DNN Model

The second proposed model shares the same architecture as the first one, with the addition of an LSTM layer at the end of the CNN part for the weather variables. Specifically, the output of the CNN part is passed to an LSTM layer consisting of 128 units. The resulting output is then combined with the output of the fully connected layer for the input _others data. This model architecture is designed to further capture the temporal dependencies and nonlinear effects of the weather variables, in addition to the high-level features extracted by the CNN part. Similar to the architecture described above, dropout layers were utilized to prevent overfitting. Specifically, four dropout layers with dropout ratios of 0.5, 0.5, 0.7, and 0.2 were respectively inserted after the CNN layer, at the LSTM layer, at the end of the fully connected layer for input _other data, and at the final layer of the model. The complete modeling architecture is illustrated in Figure 3.3.



(a) Proposed CNN-DNN Model



(b) Proposed CNN-LSTM-DNN Model

Figure 3.3: The CNN architectures proposed in this study includes convolutional, and fully connected layers denoted by Conv, and Dense respectively. The parameters of the convolutional layers are presented in the form of "convolution type— number of filters—kernel size—stride size". For all layers, "valid" padding was employed. Matrix concatenations are indicated by Ⓒ, while the symbol Ⓣ is used to indicate matrix transpose. Rectified Linear Unit (ReLU) was chosen as the activation function for all networks, with the exception of the fully connected layer in the input _other data, where a Leaky ReLU activation function was applied.

Table 3.2: Comparison of Models with and without Soil Variables

Model	No Soil Variables			All 66 Soil Variables are Included		
	RMSE	**MAE**	**r**	**RMSE**	**MAE**	**r**
Train						
RF	3.24	2.23	0.981	3.14	2.15	0.982
XGBoost	6.75	5.15	0.908	6.64	5.05	0.911
LASSO regression	8.69	6.76	0.838	8.69	6.76	0.839
GEM	-	-	-	-	-	-
Test						
RF	7.12	5.32	0.894	7.11	5.32	0.895
XGBoost	7.04	5.33	0.898	6.96	5.27	0.9
LASSO regression	9.33	7.26	0.81	9.33	7.25	0.81
GEM	6.67	5.05	0.908	6.64	5.02	0.909
Validation						
RF	7.01	5.27	0.899	7.01	5.26	0.899
XGBoost	6.98	5.32	0.901	6.87	5.22	0.904
LASSO regression	9.41	7.3	0.809	9.41	7.29	0.809
GEM	6.55	4.92	0.91	6.58	4.92	0.912

3.3.2.3 Generalized Ensemble Method

The GEM method is an advanced technique for creating a regression ensemble that combines the strengths of multiple base estimators. The method was first proposed by Perrone and Cooper [1995] in the context of artificial neural networks. The main goal of GEM is to find the optimal weights of the base models that minimize the error metric, such as MSE or RMSE. To prepare the data for model training and evaluation, we randomly partitioned the dataset into a training set containing 80% of the data (74,422 samples), a validation set including 10% of the data (9,303 samples), and a test set containing the remaining 10% of the data (9,303 samples). We selected the best performing model on the validation set and leveraged the following optimization approach to create an ensemble of models that further improved the prediction accuracy. The problem can be stated as a nonlinear convex optimization problem, where the objective is to minimize the sum of squared errors between the true values (y_i) and the predicted values ($\hat{y}_{ij}$) of all observations (i=

$1\ldots, n$) by the k base models ($j = 1,\ldots, k$). The validation set was used to optimize the ensemble weights.

$$\min_{w_j} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} w_j \hat{y}_{ij})^2 \tag{3.2}$$

The problem is subject to two constraints: the weights of all base models should be non-negative ($w_j \geq 0$) and sum up to one ($\sum_{j=1}^{k}(w_j) = 1$). Here, w_j represents the weight assigned to base model j.

3.3.3 Optimal Genotype by Environment Selection

To predict the yield for all possible genotypes in each specific location and environmental scenario, we employed the CNN-DNN model. To simplify the results, we excluded the maturity group feature and retrained the proposed model. This was necessary because different types of maturity groups were utilized in each location. The aim of this approach is to showcase the best genotypes that could be cultivated in each specific location and weather (year) scenario.

3.3.4 Design of Experiments

Since the ground truth response variables for the test data were not released after the competition, we solely relied on the the training dataset, which consisted of 93,028 observations, to train and test our proposed DL models and other ML models. The data preprocessing resulted in 6391 column features (6020 features after one-hot encoding maturity group, year, location IDs, and genotype IDs, and 371 (53×7) features after downsampling the weather data to a 4-day level). To assess the performance of the proposed GEM model and conduct comparative analyses with alternative ML models, we partitioned the dataset into training, testing, and validation sets. The training set comprises 74,422 records, the testing set contains 9,303 records, and the validation set also encompasses 9,303 records. Table 3.3 provides detailed summary statistics for each of these datasets.

Table 3.3: Summary statistics of soybean yield for the training, testing, and validation datasets. The unit of yield is bushels per acre.

Summary Statistics	**Train**	**Test**	**Validation**
Total number of locations	159	158	157
Year range	2003-2015	2003-2015	2003-2015
Mean yield	50.68	50.55	50.61
Standard deviation of yield	15.95	15.91	15.99
25th percentile of yield	39.8	39.7	
Median yield	50.70	50.50	50.50
75th percentile of yield	61.40	61.35	61.20
Minimum yield	0.40	1.80	2.70
Maximum yield	112.40	111.70	109.30
Number of weather components	7	7	7
Number of maturity groups	10	10	10
Number of genotype IDs	5,838	4,020	3,956
Number of observations	74,422	9,303	9,303

In order to make a comprehensive comparison, we incorporated three additional commonly used prediction models: RF [Breiman, 2001], XGBoost [Chen and Guestrin, 2016], and LASSO [Tibshirani, 1996]. Further details on the implementation of these models are outlined below.

•RF is an ensemble learning algorithm in machine learning. It works by combining multiple decision trees and using bagging to create a more accurate model. Bagging involves randomly sampling the dataset multiple times with replacement, creating different subsets of data for each tree to learn from. Each tree is trained on a different subset of the data, and their predictions are combined to make the final prediction. After experimenting with various numbers of trees in the RF model, we discovered that 550 trees produced the most accurate predictions. Furthermore, increasing the number of trees did not improve the accuracy but significantly increased the training time. We also examined different numbers of maximum tree depths and observed that a maximum depth of 55 generated the most precise predictions. Altering the maximum depth of the trees had a significant impact on the prediction accuracy, with an increase resulting in overfitting and a decrease leading to decreased prediction accuracy.

•XGBOOST is a popular machine learning algorithm known for its speed and accuracy in solving regression and classification problems. It is based on the concept of boosting, where weak learners are combined to form a strong learner. XGBOOST is an optimized version of gradient boosting, and it uses a tree-based model. It has several advantages, such as handling missing values, feature importance ranking, and regularization to prevent overfitting. To optimize the XGBOOST model for predicting soybean yield, we explored different hyperparameters ranges for max depth and subsample. After training and validating multiple models with different combinations of hyperparameters, we found that a max depth of 13 and a subsample of 0.7 provided the best results in terms of RMSE and MAE.

•LASSO is a linear regression technique used to analyze data with a high number of features. It uses regularization to constrain the coefficient estimates towards zero, which results in simpler models and reduces the risk of overfitting. The LASSO model adds a penalty term to the sum of the squared residuals, where the penalty is proportional to the absolute value of the coefficients. The optimization algorithm tries to minimize this penalty term along with the sum of the squared residuals. The alpha parameter in sklearn's Lasso function controls the strength of the L1 penalty on the coefficients, which is the same as the L1 term in the LASSO model. A higher alpha value will result in more coefficients being forced to zero, leading to a simpler and more interpretable model. In our study, we tried a range of alpha values, including [0.0001, 0.001, 0.01, 0.1, 1, 10, 100], and found that alpha=0.0001 provided the best result.

We maintained the same randomly partitioned dataset, which was used to train the ensemble models, consisting of a training set with 80% of the data (74,422 samples), a validation set with 10% of the data (9,303 samples), and a test set with the remaining 10% of the data (9,303 samples), for both hyperparameter tuning and model evaluation. Multiple models were trained using various hyperparameter values and their performance was evaluated on the validation set. The hyperparameter values that resulted in the best performance on the validation set were selected, and the corresponding model was evaluated on the test set to estimate its generalization performance. The range of hyperparameter values that we tested was selected based on our domain

knowledge. Table 3.4 shows the tested hyperparameters along with the best estimates obtained for the baseline models.

Table 3.4: Hyperparameters of the baseline machine learning models employed to predict soybean yield.

Model	**Parameters**	**Best Parameter**
RF	Number of estimators	550
	Max. feature numbers	Sqrt
	Max. depth	55
	Min. samples split	5
	Min. samples leaf	1
	Bootstrap	FALSE
XGBoost	Max. depth	13
	Objective	[reg:squared error]
	regularization alpha	0.0001
	Min. child weight	5
	Gamma	0.05
	Learning rate	0.09
	Booster	Gbtree
	Subsample	0.7
	Column sample by tree	0.9
LASSO regression	alpha	0.0001

The architecture and hyperparameters of the CNN-DNN and CNN-LSTM-DNN models are described in Figure 3.3. We trained the proposed models using the Adam optimizer with a scheduled learning rate of 0.0004, which decayed exponentially with a rate of 0.96 every 2500 steps. The models were trained for 800,000 iterations with a batch size of 48. ReLU was chosen as the activation function for all networks, with the exception of the fully connected layers in the input _other data and soil data, where a Leaky ReLU activation function was applied.

3.3.5 Model Evaluation

In this study, we evaluated the performance of our prediction models using two widely used metrics: MAE [Eq. 3.3] and RMSE [Eq. 4.3]. Both of these metrics provide a measure of the distance between the predicted and actual values of the target variable. Specifically, MAE

represents the average absolute difference between the predicted and actual values, while RMSE represents the square root of the average of the squared differences between the predicted and actual values. By using both of these metrics, we were able to assess the accuracy of our models and compare their performance against each other. We also reported the results of correlation coefficient (r) [Eq. 3.5] as an additional metric to evaluate the linear relationship between the predicted and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.4}$$

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3.5}$$

Where n is the total number of data points, y_i is the true value of the i-th data point, $\hat{y}_i$ is the predicted value of the i-th data point, and $\bar{y}$ and $\bar{\hat{y}}$ represent their respective means.

3.4 Results

Our proposed GEM model achieved an impressive RMSE of 5.95 and MAE of 4.47 on the test set, earning us third place in the competition. In this paper we did one step further and used our proposed CNN-DNN model to select the top 10 optimal genotypes with the highest yields in each specific location and environmental scenario. Since the ground truth response variables for the test data were not released after the competition, we only used the training dataset (93,028 samples) for yield prediction and further analysis. Specifically, we used the training dataset to select the top 10 optimal genotypes with the highest yields in each specific location and environmental scenario. In this section, we will first examine the results of the yield prediction, followed by the selection of the

top 10 optimal genotypes that yielded the highest yields for each particular location and environmental condition.

3.4.1 Prediction results

The study employs the best hyperparameter settings obtained through hyperparameter tuning to train and validate three machine learning models, and our proposed hybrid deep learning models. The performance of the baseline machine learning models and the proposed GEM model in predicting soybean yield is evaluated based on the test and validation results of RMSE, MAE, and r (correlation coefficient), and it is presented in Table 3.5.

Table 3.5: Comparison of Test and Validation Results of RMSE, MAE, and r for the Baseline Machine Learning Models and Proposed GEM Model in Predicting Soybean Yield.

	RF	**XGBoost**	**LASSO regression**	**Proposed GEM**
Train RMSE	3.24	6.75	8.69	-
Test RMSE	7.12	7.04	9.33	6.67
Validation RMSE	7.01	6.98	9.41	6.55
Train MAE	2.23	5.15	6.76	-
Test MAE	5.32	5.33	7.26	5.05
Validation MAE	5.27	5.32	7.30	4.92
Train r	0.981	0.908	0.838	-
Test r	0.894	0.898	0.810	0.908
Validation r	0.899	0.901	0.809	0.91

The performance of the proposed GEM model, which combines the CNN-DNN model and the CNN-LSTM-DNN model, was compared with several other machine learning models including XGBoost, RF, and LASSO. The results showed that the GEM model outperformed all other tested models. The reason for the outperformance can be attributed to the GEM model's ability to capture the nonlinearity of weather data and its capacity to capture the temporal dependencies of weather data.

While XGBoost and RF are powerful machine learning models, they rely heavily on linear relationships and may not be able to capture the nonlinear relationships present in the weather data. LASSO, on the other hand, is a linear regression model with an L1 penalty, which can result

in some of the coefficients being forced to zero. While this can result in a simpler and more interpretable model, it may not be able to capture the complex relationships present in the weather data, and genotype by environment interactions.

The GEM model, on the other hand, combines the strengths of multiple models, including the highly nonlinear structure of the CNN model and the ability to capture the temporal dependencies of weather data using the LSTM model. This results in a more robust and accurate model that outperforms the other tested models.

The hexagonal plots shown in Figure 3.4 are a visualization tool used to compare the ground truth yield with the predicted yield values for different machine learning models. The plots show the density of points where the two yields overlap, with the color of the hexagons representing the density of points. The 1:1 line represents the ideal situation where the predicted yield is exactly equal to the ground truth yield.

By looking at the hexagonal plots and the position of the points relative to the 1:1 line, we can observe how well each model is performing. If the points are concentrated near the 1:1 line, it indicates that the model is performing well, with high accuracy and precision. On the other hand, if the points are scattered or far from the 1:1 line, it indicates that the model is not performing well and is making large errors in its predictions.

In this case, based on the hexagonal plots in Figure 3.4, the GEM model has the most tightly clustered predicted yield values around the 1:1 line, suggesting that it is the most accurate model in predicting soybean yield. The RF model has a slightly wider spread of predicted yield values, indicating slightly less accuracy. The XGboost model also shows a strong positive correlation with the ground truth yield values, but has more scatter than the RF model. The performance of the LASSO model was comparatively weaker, which was demonstrated by the scattered data points that exhibited more deviation from the 1:1 line compared to other models.

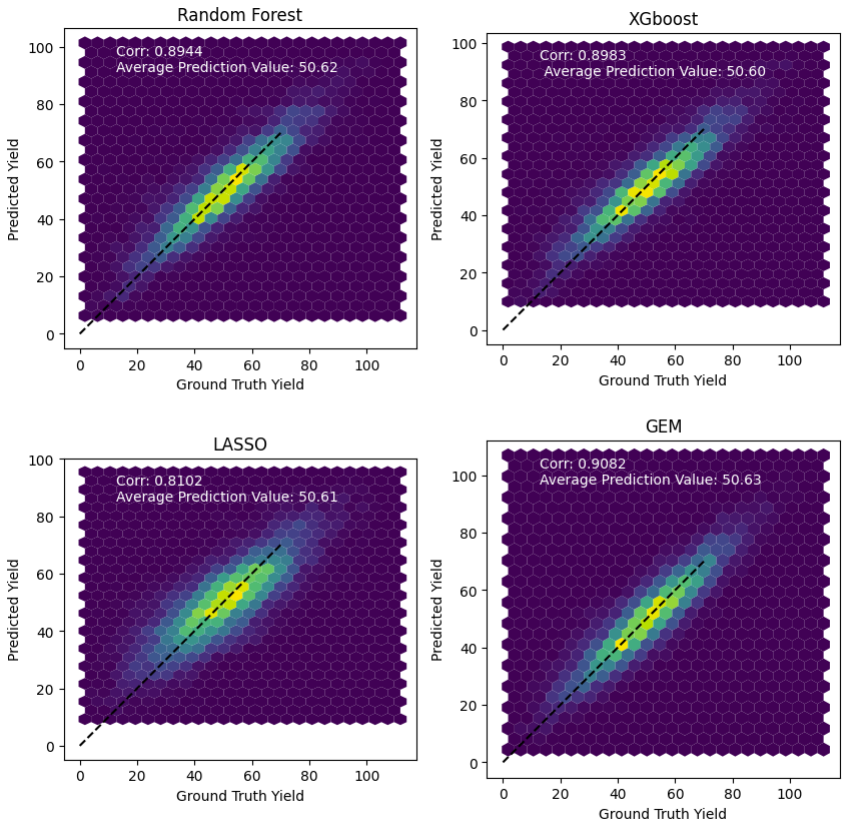


Figure 3.4: Hexagonal plots of the predicted soybean yield vs. ground truth yield values for the three machine learning models and proposed GEM model on the test data.

Figure 3.5 illustrates the spatial distribution of average prediction errors for soybean yield in test data using the proposed GEM, RF, and XGBoost models, and average observed yield values across 28 U.S. states and Canadian provinces. This figure allows for the identification of states/provinces with higher average error percentages, providing valuable insights to enhance data collection in those regions. The GEM model reveals a variation in average error percentages ranging from 8.19% to 44.68% across 28 U.S. states and Canadian provinces in the test data. As anticipated, states and provinces such as Texas and Manitoba, which have a lower number of observations as depicted in Figure 3.2, exhibit higher prediction errors. Although the proposed GEM model

exhibits a wider range of average percentage error values compared to the RF and XGBoost models, as indicated in Table 3.5, the overall performance of the GEM model surpasses that of the baseline models. The prediction percentage error for location i is determined using Equation 3.6, where the difference between the predicted yield value and the actual yield value for each location within each state or province is divided by the actual yield value. We then averaged the prediction errors of all locations within each state or province and displayed the results in Figure 3.5.

$$\text{Prediction Error Percentage}_i = \frac{\text{Actual Yield}_i - \text{Predicted Yield}_i}{\text{Actual Yield}_i} \times 100 \tag{3.6}$$

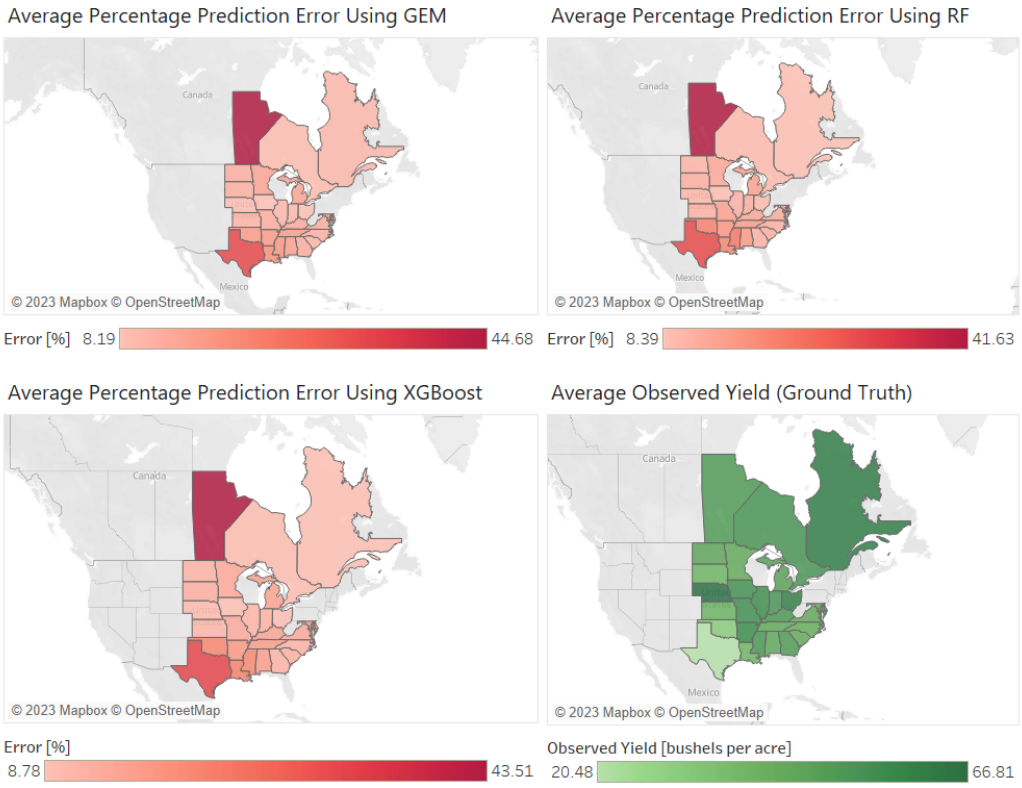


Figure 3.5: Spatial distribution of average prediction errors for soybean yield in test data using the proposed GEM, RF, and XGBoost models, and average observed yield values across 28 U.S. states and Canadian provinces.

3.4.2 Optimal genotype selection

In this section, we utilized the entire dataset to identify the top 10 genotypes with the highest yields for each location-environment combination. Based on the results from the previous section, the GEM model indicated that the CNN-DNN model had the highest weight. Therefore, we proceeded to retrain the model on the entire dataset, excluding the maturity group. Following this, the model was used to predict yields for all 5,838 genotypes across different weather and location combinations. Next, we chose the top 10 genotypes with the highest yields. We then proceeded to compute the average yield for these elite genotypes across each location-environment combination. However, due to the unavailability of weather data for all years and locations, we ended up with varying amounts of weather data for each location. For instance, for location ID 167, we selected the top 10 genotypes with the highest yields for each of weather variables from 2008 to 2010, whereas for location ID 163, we selected the top 10 genotypes with the highest yields for each weather variables from 2004 to 2012.

To illustrate the quality of these selected genotypes, we created a Tableau Public visualization. This visualization showcases the difference between the average predicted yield of optimal genotypes and the actual yield of existing genotypes across all locations in each state and year. This visualization is accessible at the following link: Tableau Public Visualization

Within this Tableau page, the option is available to select a specific year and examine the differences between the average predicted yield of optimal genotypes and the actual yield of existing genotypes across all locations in each state for that year. The range of differences observed across all years suggests that the optimal genotypes can potentially lead to increased average soybean yields in all states, with differences ranging from at least 5.1 to 42.5 bushels per acre.

These visualizations underscore the nuanced influence of weather conditions on the selection of optimal genotypes for achieving the highest yields. Moreover, they emphasize the critical role of genotype choice in varying weather conditions. For instance, consider Location ID 1, located in the state of Louisiana (LA). Depending on the weather variables corresponding to the year, the top 10 genotypes for the highest predicted yields varied, exemplifying the sensitivity of optimal genotype

selection to different weather conditions. Significantly, the impact of weather variables on achieving the highest yield with optimal genotypes is magnified when considering diverse Location IDs, states, and provinces. The variability in weather variables for different years also contributes to the observed differences in yield outcomes. Notably, the absence of weather variables for certain years in specific location IDs adds another layer of complexity. Illustratively, for weather variables corresponding to the year 2004, the range of differences spans from 6.04 for the state of DE (Delaware) to 42.5 for the province of MB (Manitoba). In contrast, for weather variables corresponding to the year 2010, the range narrows to 11.05 for the state of IA (Iowa) and 23.73 for the province of MB (Manitoba). The highest differences between the average predicted yields of optimal genotypes and the actual yields of existing genotypes are inherently linked to the specific weather variables corresponding to different years. It is noteworthy that the years 2014 and 2015 exhibit a lower number of locations in our analysis.

3.5 Analysis

3.5.1 Feature importance analysis using RMSE change

In this study, we conducted a feature importance analysis to identify the key predictors that significantly influence our model's predictions. The analysis is based on the RMSE change, which measures the impact of feature permutations on prediction performance. This method allowed us to assess the impact of variable shuffling on the model's performance.

• **Baseline RMSE Calculation**: We initially computed the baseline RMSE ($r0$) using the proposed GEM model predictions ($yhat$) and the ground truth values (test set containing the remaining 10% of the data (9,303 samples)).

• **Permutation and RMSE Change**: We systematically shuffled the columns within various groups of variables and recalculated the RMSE for each permutation. These groups encompassed variables related to weather conditions, such as ADNI, AP, ARH, MDNI, MaxSur, MinSur, and AvgSur. Additionally, we considered other critical variables, including MG, year, location, and genotype ID. Among these, the categorical variables, such as MG, year, location, and genotype IDs,

underwent one-hot encoding, resulting in multiple variables representing these categories. Similarly, each weather-related variable comprises 53 distinct variables, each signifying the aggregation of daily feature values through the process of averaging and downscaling the data to a 4-day granularity.

• **Interpreting RMSE Change**: A higher RMSE change after shuffling indicates that the original group of variables had a more substantial impact on the model's predictions. In other words, when these variables are shuffled, the model's performance degrades significantly because they were contributing significantly to the model's accuracy.

Conversely, a lower RMSE change after shuffling suggests that the original group of variables had a lesser influence on the model's predictions. Shuffling these variables doesn't significantly impact the model's performance, indicating that they might not be as critical for prediction accuracy.

Figure 3.6 illustrates the RMSE changes for different groups of variables after shuffling. Each group represents a set of variables, and the RMSE change quantifies the impact of shuffling those variables on the model's predictions. Each group in the analysis signifies a distinct set of variables.

Among all the groups, the location variable exhibits the highest RMSE change, suggesting that it plays a pivotal role in the model's predictions. When the location variable is shuffled, there is a significant decline in model performance. This emphasizes the substantial influence of geographical location on prediction accuracy. Following location, the MG variable shows the second-highest RMSE change. Shuffling this variable results in a noticeable reduction in model accuracy. This highlights the significance of considering the maturity stage of crops or plants for accurate predictions. Different maturity groups of soybeans have varying growth and flowering patterns, affecting the timing of yield. The year variable ranks third in terms of RMSE change. Shuffling the year variable leads to a substantial drop in model performance. This implies that variations across different years significantly affect the model's ability to make accurate predictions, likely due to year-specific climate patterns or other time-dependent factors. The genotype variable occupies the fourth position in RMSE change. Its shuffling causes a notable decrease in model accuracy, underscoring its importance in achieving reliable predictions. Different genotypes or plant varieties

evidently contribute significantly to the model's predictive power. Within the weather category, MDNI demonstrates the highest RMSE change, followed by AP, MinSur, ADNI, AvgSur, MaxSur, and ARH. While these weather-related variables do influence the model's predictions, their impact appears to be less pronounced compared to the location, MG, year, and genotype variables. Shuffling these weather variables results in a relatively modest effect on model performance, suggesting that they may be less critical for prediction accuracy compared to the aforementioned groups.

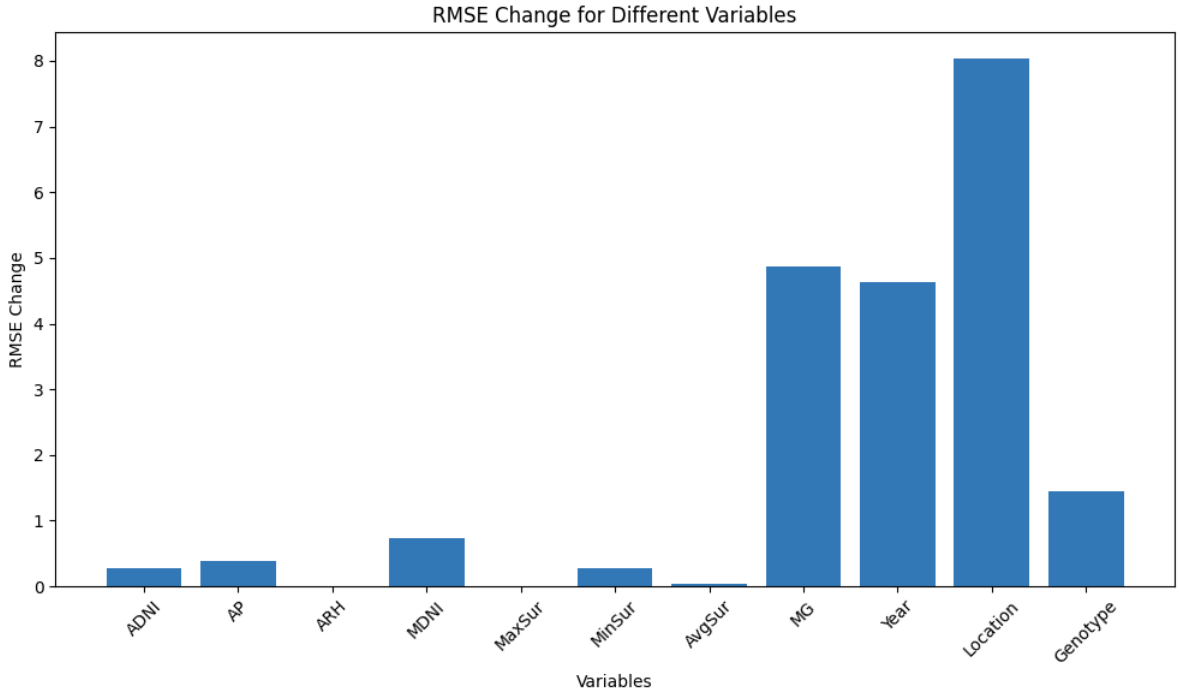


Figure 3.6: The RMSE changes for different groups of variables after shuffling. Each group represents a set of variables, and the RMSE change quantifies the impact of shuffling those variables on the model's predictions.

In our analysis, we observed that certain time periods within the weather variables exhibited the highest RMSE change after shuffling. Specifically, for two key weather variables, Maximum Direct Normal Irradiance (MDNI) and Average Precipitation (AP), we identified the time periods that demonstrated the most significant impact on model performance.

For MDNI, we found that time period 25, which corresponds to approximately week 15^{th}, exhibited the highest RMSE change following shuffling, as it is shown in Figure 3.7. Similarly, for AP, time period 29 (approximately week 17^{th}) showed the highest RMSE change, as illustrated in Figure 3.8. These findings prompt us to explore the relationship between these time periods and the growth stages of soybeans in the United States.

In the context of soybean growth in the U.S., the growth stages are often categorized into Vegetative (V) and Reproductive (R) stages. Based on our analysis and considering typical soybean growth patterns in the USA McWilliams et al. [1999], University of Kentucky Cooperative Extension [nd], we can provide the following insights:

• Week 10 (Approximately): During this time, soybeans are in the early to mid-vegetative stages, typically ranging from V4 to V6. They are transitioning from early vegetative growth to the onset of reproductive growth University of Kentucky Cooperative Extension [nd].

• Week 12 (Approximately): At this stage (mid to late June), soybeans are typically in the V6 to V8 vegetative stage, indicating that they are approaching the reproductive stages University of Kentucky Cooperative Extension [nd].

• Week 15 (Approximately): This period, occurring in early to mid-July, corresponds to soybeans being in the V8 to V10 vegetative stage. This is a critical time when soybeans start transitioning to early reproductive stages, with some plants beginning to flower (R1 stage) University of Kentucky Cooperative Extension [nd].

• Week 17 (Approximately): Around mid to late July, soybeans may have progressed to the R2 (Full Flower) to R3 (Beginning Pod) stages. This is a vital phase during which soybeans flower and initiate pod development University of Kentucky Cooperative Extension [nd].

The observed highest RMSE changes in time periods 25 (MDNI) and 29 (AP) suggest a noteworthy correlation with soybean growth stages. The RMSE changes in these weeks signify the sensitivity of soybean growth to solar radiation (MDNI) and precipitation (AP) during key reproductive and pod development stages. These findings underscore the significance of weather variables during crucial growth phases of soybeans and their influence on accurate yield predictions.

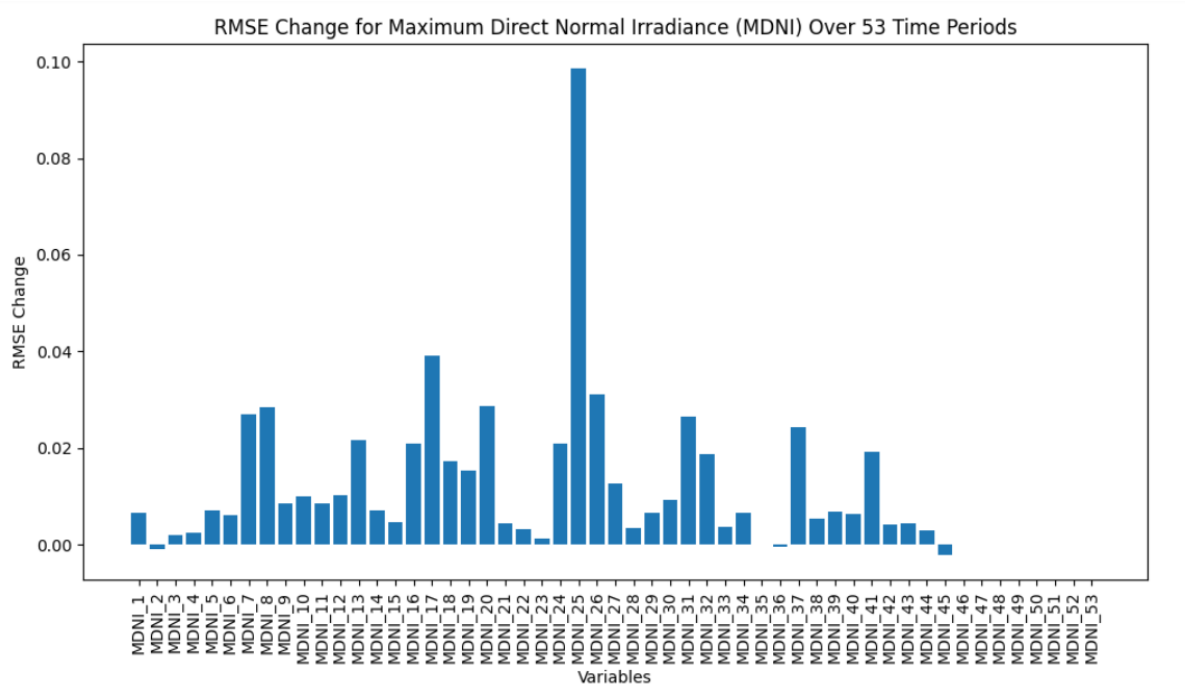


Figure 3.7: The RMSE changes resulting from the systematic shuffling of Maximum Direct Normal Irradiance (MDNI) variables. The MDNI variables represent 4-day intervals throughout the growing season, spanning from the first to the 53rd interval, and each bar corresponds to the RMSE change associated with shuffling a specific MDNI variable.

3.5.2 Impact of state-level soil characteristics on Soybean yield prediction

As highlighted in Section 3.2, the dataset lacks information about soil variables for each location. The available data only includes location IDs and states, with no provided latitude and longitude details. Consequently, the exact geographical coordinates for each location ID are unavailable, limiting our spatial information to state-level knowledge. In this section, we aim to investigate the impact of incorporating soil data in addition to the variables provided by the MLCAS2021 Crop Yield Prediction Challenge. To enhance our dataset with soil information, we utilized preprocessed and cleaned soil data available from an open-source repository on Github. The soil data originates from SoilGrids250m and comprises 11 variables measured at six different

depths (0-5cm, 5-15cm, 15-30cm, 30-60cm, 60-100cm, 100-200cm) with a resolution of 250 square meters. The corresponding acronyms and properties of these soil variables are listed in Table 3.6.

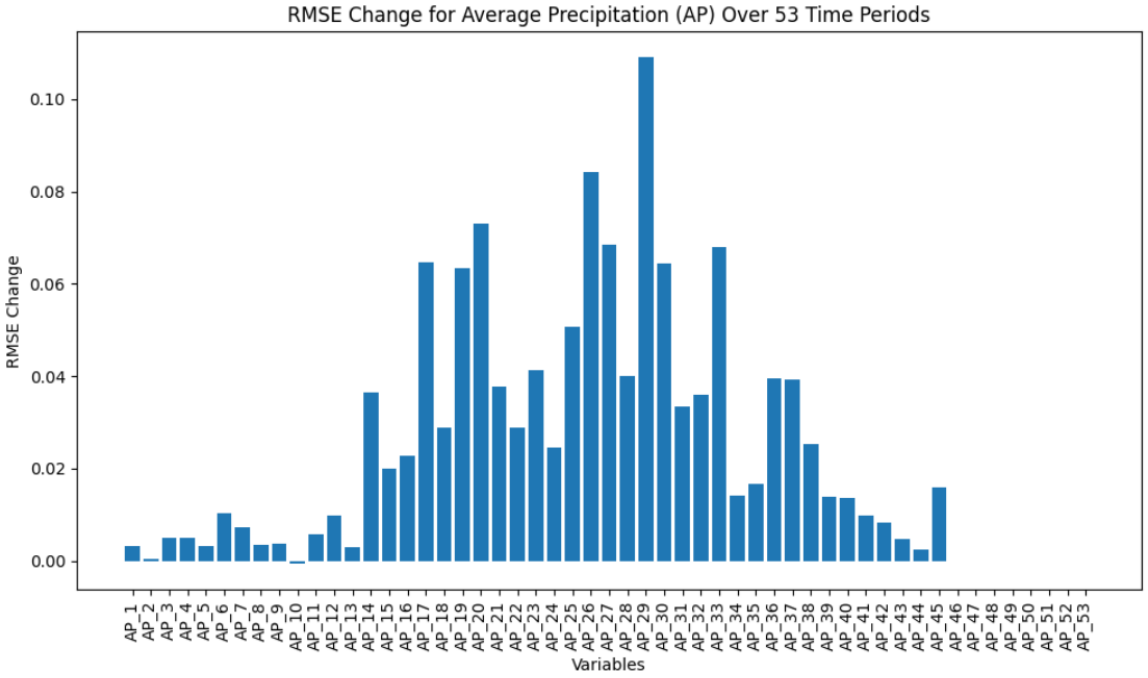


Figure 3.8: The RMSE changes resulting from the systematic shuffling of Average Precipitation (AP) variables. The AP variables represent 4-day intervals throughout the growing season, spanning from the first to the 53rd interval, and each bar corresponds to the RMSE change associated with shuffling a specific AP variable.

To understand how soil data affects our analysis, we added 66 soil variables to each record in our dataset. We merged the soil data with our existing dataset using the State column. This ensures that all locations within the same state share the same soil information. The employed CNN-DNN and CNN-DNN-LSTM models, detailed in Sections 3.3.2.1 and 3.3.2.2 respectively, remained consistent with the methodology outlined in this paper. The key modification involved integrating soil data via an additional dense layer. This layer, structured with 512 neurons, processes the input soil data using a Leaky ReLU activation function. To prevent overfitting and enhance the model's robustness, dropout regularization is applied. This regularization technique

Table 3.6: Acronyms and Corresponding Soil Properties

Acronym	**Property**
bdod	Bulk density
cec	Cation exchange capacity at pH7
cfvo	Coarse fragments
clay	Clay
nitrogen	Total Nitrogen
ocd	Organic carbon density
ocs	Organic carbon stock
phh2o	pH in H2O
sand	Sand
silt	Silt
soc	Soil organic carbon

randomly drops out approximately 50% of the neurons during training, preventing the model from relying too heavily on specific features and aiding in better generalization. The architectural configurations of both the CNN-DNN and CNN-LSTM-DNN models, incorporating soil data, are visually represented in Figure 3.9.

We maintained consistency in our experimental setup by employing identical training, test, and validation datasets, along with the same set of hyperparameters, for both traditional ML models and our proposed GEM model, as outlined in Subsection 3.3.4. Additionally, after incorporating soil variables into the existing data, we conducted a comparison of the models' performances. The results of this comparison are presented in Table 3.7, shedding light on the impact of integrating soil variables into the predictive models. Given the lack of precise latitude and longitude data for each location, we resorted to using state-level soil data. Therefore, the inclusion of soil variables as additional input features did not yield substantial improvements in model performance, which aligns with our expectations. Noteworthy is the marginal impact observed in the RMSE for the Test data. Specifically, the RMSE decreased by only 0.14%, 1.14%, and 0.44% for the RF, XGBoost, and GEM models, respectively. LASSO regression exhibited no change in RMSE for the Test data. This limited improvement can be attributed to the inherent challenges in leveraging soil

data. Consequently, despite our efforts, the added granularity from soil data did not significantly enhance model performance.

Table 3.7: Comparison of Models with and without Soil Variables

Model	**No Soil Variables**			**All 66 Soil Variables are Included**		
	RMSE	**MAE**	**r**	**RMSE**	**MAE**	**r**
Train						
RF	3.24	2.23	0.981	3.14	2.15	0.982
XGBoost	6.75	5.15	0.908	6.64	5.05	0.911
LASSO regression	8.69	6.76	0.838	8.69	6.76	0.839
GEM	-	-	-	-	-	-
Test						
RF	7.12	5.32	0.894	7.11	5.32	0.895
XGBoost	7.04	5.33	0.898	6.96	5.27	0.9
LASSO regression	9.33	7.26	0.81	9.33	7.25	0.81
GEM	6.67	5.05	0.908	6.64	5.02	0.909
Validation						
RF	7.01	5.27	0.899	7.01	5.26	0.899
XGBoost	6.98	5.32	0.901	6.87	5.22	0.904
LASSO regression	9.41	7.3	0.809	9.41	7.29	0.809
GEM	6.55	4.92	0.91	6.58	4.92	0.912

3.6 Conclusion

In this study, we proposed two novel CNN architectures that incorporate a 1-D convolution operation and an LSTM layer. These models were developed to predict soybean yield using a combination of factors, including maturity group, genotype ID, year, location, and weather data. Our study is based on an extensive dataset collected from 159 locations across 28 U.S. states and Canadian provinces over a span of 13 years. These architectures represent a significant advancement in the field of crop yield prediction, allowing us to leverage the power of deep learning to improve accuracy and efficiency in genotype selection. Moreover, we have employed the GEM method to

determine the optimal weights of our proposed CNN-based models, which has led to superior performance in the MLCAS2021 Crop Yield Prediction Challenge and compared to baseline models.

Our work has gone beyond traditional crop yield prediction methods by addressing the challenge of genotype by environment interaction, which is a critical factor in selecting genotypes for increased crop yields, particularly in the face of global climate change. Conventionally, plant breeders rely on extensive field testing of hybrids to identify those with the highest yield potential, a process that is both time-consuming and resource-intensive. Our approach has introduced a data-driven paradigm for genotype selection, wherein we use environmental data and genotype information to predict crop yields. This approach enables us to identify the most efficient genotypes for each location and environmental condition by forecasting crop yields based on weather conditions and then selecting the optimal genotype with the highest yield. This novel strategy holds the potential to significantly enhance policy and agricultural decision-making, optimize production, and ensure food security.

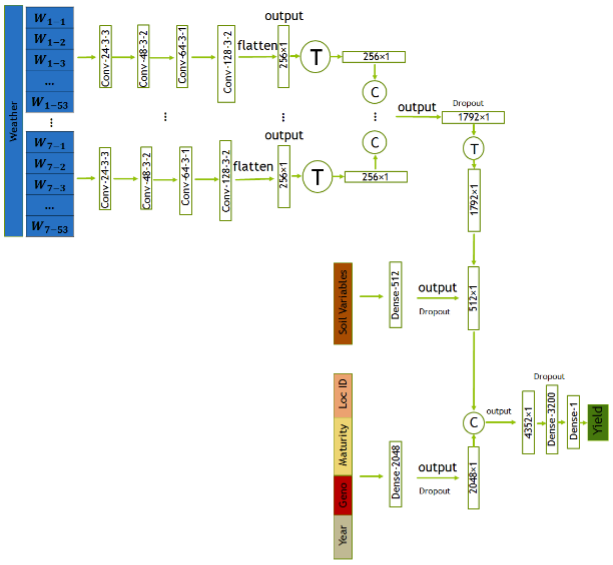
In our analysis, we evaluated our proposed GEM model against three commonly used prediction models: RF, XGBoost, and LASSO. The GEM model demonstrated notable performance advantages across several metrics. Specifically, it exhibited lower RMSE and MAE values ranging from 5.55% to 39.88% and 5.34% to 43.76%, respectively, compared to the baseline models when evaluated on test data. Additionally, the GEM model showcased higher correlation coefficients ranging from 1.1% to 10.79% in comparison to the baseline models. These performance improvements suggest the effectiveness of the GEM model in soybean yield prediction, attributed to its ability to capture the nonlinear nature of weather data and model the temporal dependencies of weather variables, including genotype by environment interactions. This is achieved through the combination of two CNN-based models, which are adept at handling complex relationships in the data.

Additionally, we conducted a feature importance analysis using RMSE change to identify significant predictors affecting the model's predictions. The location variable had the highest RMSE change, indicating its strong influence on predictions. MG, year, and genotype also played

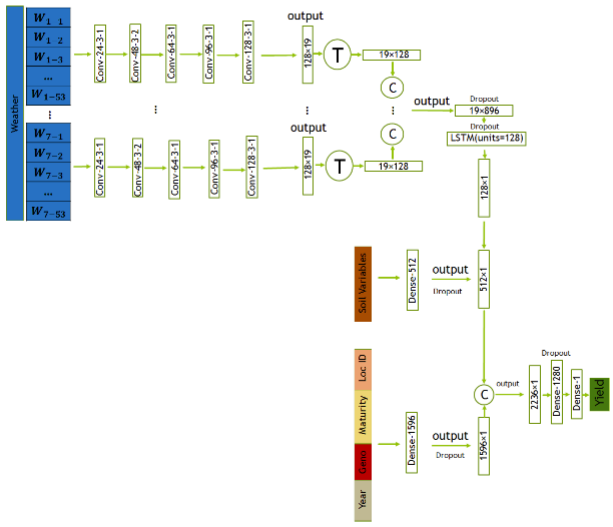
crucial roles. Among weather variables, Maximum Direct Normal Irradiance (MDNI) had the most impact, followed by Average Precipitation (AP), Minimum Surface Temperature (MinSur), and others. While weather variables influenced predictions, categorical variables like location and MG were more influential.

In addition to assessing the importance of different variable groups, we delved deeper into the temporal aspects of weather data. Specifically, we investigated significant time periods within MDNI and AP variables that exhibited the highest RMSE change after shuffling. The highest RMSE changes observed in time periods 25 (week 15^{th}) (MDNI), and 29 (week 17^{th}) (AP) point to a significant link with soybean growth stages. These RMSE fluctuations in these weeks highlight how soybean growth is affected by solar radiation (MDNI) and precipitation (AP) during important reproductive and pod development stages.

Despite the constraints imposed by limited information and the absence of exact latitudes and longitudes in our dataset, we opted to explore the impact of soil variables on model performance. We accommodated this limitation by integrating state-level soil variables into the original dataset. Our findings suggest that the integration of soil variables, under the current data constraints, did not lead to a substantial enhancement in the predictive capabilities of the models. Given that climate change can also have an adverse effects on soil attributes [Das et al., 2016], it is advisable to consider datasets that provide precise soil variables for each specific location. This more granular data can significantly enhance our understanding of the intricate relationships among weather, soil, and crop outcomes. The incorporation of location-specific soil attributes into predictive models has the potential to elevate accuracy, particularly in regions where soil quality plays a pivotal role in agricultural outcomes.



(a) Proposed CNN-DNN Model Incorporating Soil Data



(b) Proposed CNN-LSTM-DNN Model Incorporating Soil Data

Figure 3.9: The CNN architectures proposed in this study includes convolutional, and fully connected layers denoted by Conv, and Dense respectively. The parameters of the convolutional layers are presented in the form of "convolution type— number of filters—kernel size—stride size". For all layers, "valid" padding was employed. Matrix concatenations are indicated by Ⓒ, while the symbol Ⓣ is used to indicate matrix transpose. Rectified Linear Unit (ReLU) was chosen as the activation function for all networks, with the exception of the fully connected layers in the input _other data and soil data, where a Leaky ReLU activation function was applied.

3.7 Acknowledgments

We express our gratitude to the organizers of the Third International Workshop on Machine Learning for Cyber-Agricultural Systems (MLCAS 2021) for their diligent efforts in arranging the MLCAS2021 Crop Yield Prediction Challenge and generously sharing the invaluable datasets. It is noteworthy that this manuscript has also been made available as a preprint on arXiv. This work was partially supported by NSF and USDA (#1830478, #1842097, and #2021-67021-35329).

3.8 References

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee.

Arzanipour, A. and Olafsson, S. (2022). Evaluating imputation in a two-way table of means for training data construction. *SSRN*.

Bertan, I., de Carvalho, F. I., and Oliveira, A. C. d. (2007). Parental selection strategies in plant breeding programs. *Journal of crop science and biotechnology*, 10(4):211–222.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., and Liew, X. Y. (2021). Automated agriculture commodity price prediction system with machine learning techniques. *arXiv preprint arXiv:2106.12747*.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13(11):114003.

Das, I., Dutta, D., and Rakshit, A. (2016). Potential effects of climate change on soil properties: A review.

de Los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nature communications*, 11(1):1–10.

Domingues, T., Brandão, T., and Ferreira, J. C. (2022). Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. *Agriculture*, 12(9):1350.

Dorffner, G. (1996). Neural networks for time series processing. *Neural network world*, 6(4):447–468.

Gangopadhyay, T., Tan, S. Y., Huang, G., and Sarkar, S. (2018). Temporal attention and stacked lstms for multivariate time series prediction.

Hajjarpoor, A., Nelson, W. C., and Vadez, V. (2022). How process-based modeling can help plant breeding deal with g x e x m interactions. *Field Crops Research*, 283:108554.

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and applied genetics*, 127(2):463–480.

Hochreiter, S. and Schmidhuber, J. (1996). Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, 9.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., and Anderson, M. (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the us midwest. *Environmental Research Letters*, 15(6):064005.

Khaki, S., Khalilzadeh, Z., and Wang, L. (2020a). Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach. *Plos one*, 15(5):e0233382.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621.

Khaki, S., Wang, L., and Archontoulis, S. V. (2020b). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750.

Kumar, M. (2016). Impact of climate change on crop yield and role of model for achieving food security. *Environmental Monitoring and Assessment*, 188(8):1–14.

Li, X., Zhou, Z., Chen, L., and Gao, L. (2019). Residual attention-based lstm for video captioning. *World Wide Web*, 22:621–636.

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Lowe, M., Qin, R., and Mao, X. (2022). A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. *Water*, 14(9):1384.

McWilliams, D. A., Berglund, D. R., and Endres, G. (1999). Soybean growth and management quick guide.

MLCAS, 2021. MLCAS2021 Crop Yield Prediction Challenge.

Nations, U., of Economic, D., and Social Affairs, P. D. (2017). World population prospects: the 2017 revision, key findings and advance tables. *Working Paper No. ESA/P/WP/248 ed*.

Oikonomidis, A., Catal, C., and Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, 36(1):2031822.

Patil, S. S. and Thorat, S. A. (2016). Early detection of grapes diseases using machine learning and iot. In *2016 second international conference on Cognitive Computing and Information Processing (CCIP)*, pages 1–5. IEEE.

Perrone, M. P. and Cooper, L. N. (1995). When networks disagree: Ensemble methods for hybrid neural networks. In *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*, pages 342–358. World Scientific.

Roberts, M. J., Braun, N. O., Sinclair, T. R., Lobell, D. B., and Schlenker, W. (2017). Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters*, 12(9):095010.

Shahhosseini, M., Hu, G., Khaki, S., and Archontoulis, S. V. (2021). Corn yield prediction with ensemble cnn-dnn. *Frontiers in plant science*, 12:709008.

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, 16(6):e0252402.

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., and Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*, 12(1):3215.

Sun, J., Di, L., Sun, Z., Shen, Y., and Lai, Z. (2019). County-level soybean yield prediction using deep cnn-lstm model. *Sensors*, 19(20):4363.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

University of Kentucky Cooperative Extension (n.d.). Soybean growth and development. Accessed: 9/13/2023.

Veenadhari, S., Mishra, B., and Singh, C. (2011). Soybean productivity modelling using decision tree algorithms. *International Journal of Computer Applications*, 27(7):11–15.

Xie, Y., Liang, R., Liang, Z., and Zhao, L. (2019). Attention-based dense lstm for speech emotion recognition. *IEICE TRANSACTIONS on Information and Systems*, 102(7):1426–1429.

Xiong, W., Reynolds, M., and Xu, Y. (2022). Climate change challenges plant breeding. *Current Opinion in Plant Biology*, 70:102308.

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M. S., Varshney, R. K., Prasanna, B. M., and Qian, Q. (2022). Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. *Molecular Plant*.

CHAPTER 4. COMPREHENSIVE CROP YIELD PREDICTION USING TRANSFORMER-ENHANCED NEURAL NETWORKS CONSIDERING DIFFERENT COMBINATIONS OF SEQUENTIAL DATA INCLUDING WEATHER, GENOTYPE, AND APSIM DATASETS AND NON-SEQUENTIAL DATA

Zahra Khalilzadeh[1], Saiara Samira Sajid[2], Saeed Khaki[3], Lizhi Wang[4], and Guiping Hu[5]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[2] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[3] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50021, USA

[4]School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74078, USA

[5]School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74078, USA

Modified from a manuscript to be submitted to *PLOS One*

4.1 Abstract

This study investigates the effectiveness of integrating transformer models with fully connected (FC) neural networks to enhance corn yield predictions. Our objectives encompass evaluating the performance of a hybrid deep learning approach that combines transformer models for sequential data with FC neural networks; investigating which combinations of phenotypic, genotypic, soil, weather, environmental covariate data derived from an unpublished APSIM crop model (APSIM

data), and metadata provide the most accurate predictions; and assessing the robustness of our proposed Transformer-Enhanced Neural Networks models by extending our analysis to incorporate temporal, genomic, and geographic extrapolations, thereby evaluating its performance across diverse variable combinations. The study employs a systematic merging process across various stages to evaluate the effects of different variable combinations. Initially, all available datasets, including genotype, trait data, metadata, soil data, weather data, and APSIM data, are merged. Subsequent stages involve excluding specific datasets to create distinct combinations. This process yields eight unique combinations of datasets, facilitating the development of models that strategically leveraged temporal dependencies in weather data, spatial and temporal correlations in APSIM data, and genetic linkages among adjacent genetic markers. Four sets of two-layer transformer models are designed to handle weather, APSIM variables including EC phenological period-soil layer and EC-phenological period, and genotype features. The input_other data, which includes trait data, metadata, and the option to include or exclude soil data, is used as input for the fully connected neural networks. Consequently, eight distinct model configurations are crafted, each exploring various combinations of variables and model architectures. These models are systematically evaluated to discern their efficacy in predicting corn yield across different variable compositions. The pre-2021 data form our training set, while the 2021 data serve as our test set. To comprehensively evaluate our proposed model, we compare its performance against six machine learning models across different variable combinations. Our results demonstrate the superiority of the Transformer-Enhanced models across all variable combinations, highlighting its capability to handle sequential data effectively. Furthermore, our analysis reveals the impact of dataset composition on model performance, with the variable combination that include weather and genotype data and exclude APSIM and soil datasets showing the most accurate prediction. Additionally, in our study, we compare transformer models with one-dimensional convolutional neural networks (CNNs) to assess their performance in handling sequential data. Our analysis reveals that the proposed Transformer-Enhanced models excel in handling sequential data. Particularly noteworthy is the significant decrease in RMSE observed for variable combination 5

(VC5), encompassing all datasets except genotype data, and subsequently for VC1, comprising all datasets, with reductions of 44% and 32%, respectively. In our study, we extend our analysis to encompass temporal, genomic, and geographic extrapolations, aiming to evaluate the robustness of our proposed Transformer-Enhanced Neural Networks models across various variable combinations in crop yield prediction. These datasets exclude records containing unique hybrids and field locations present in the test datasets, which include data from the year 2021, for the first four variable combinations involving genotype information. Subsequently, the models are tested on the same test datasets as before, with the exclusion of specific variables denoting years, hybrid names, and locations. The results suggest that the prediction error of our proposed Transformer-Enhanced Neural Networks models did not significantly increase compared to the corresponding results focusing solely on temporal extrapolation. This indicates that our proposed Transformer-Enhanced Neural Networks models effectively generalize yield predictions to untested years, hybrids, and locations.

4.2 Introduction

Accurate prediction of crop yield holds significant advantages for global food production. It facilitates informed import and export decisions crucial for national food security, empowers farmers to make knowledgeable management choices, and enhances the efficiency of the overall food supply chain [Khaki and Wang, 2019, Jame and Cutforth, 1996, Horie et al., 1992]. However, predicting crop yield can be highly challenging, given the reliance on numerous intricate factors. For example, genotype information is often characterized by numerous genetic markers, each contributing minimally to the overall variance. Identifying crucial genetic markers and estimating their effects among the vast number of markers, ranging from thousands to millions, poses a considerable challenge. Furthermore, the impact of genetic markers may involve interactions with other factors, such as environmental conditions and management practices.

There have been many attempts to represent the phenotype (such as yield) as an explicit function of the genotype (G), the environment (E), and their interactions ($G \times E$). Some of the

earliest methods ignored the $G \times E$ interaction and just considered the additive effects of G and E, letting their interactions be treated as noise [DeLacy et al., 1996, Heslot et al., 2014]. An alternative method to study $G \times E$ is to divide the environment into some mega-environments to decrease the $G \times E$ within the mega-environment[Heslot et al., 2014]. For instance, Gauch et al. used AMMI to group environments and considered additive components for G and E as main effects and multiplicative components for $G \times E$ [Gauch Jr, 2006, Hongyu et al., 2014, Sa'diyah and Hadi, 2016]. Cooper and DeLacy used agglomerative hierarchical clustering to group environments [Cooper and DeLacy, 1994]. Some studies used factorial regression to predict $G \times E$ by identifying environmental components responsible for $G \times E$ and determining the amount of genotype sensitivity to these components [Piepho, 1998, Denis, 1988]. Crop models, sets of equations determined by a few genotypes affected by various environmental conditions, have also been used for analyzing $G \times E$ [Heslot et al., 2014, Messina et al., 2009]. However, crop models do not consider many genetic variations [Hammer et al., 2002]. Linear mixed models have been used to study $G \times E$. Montesinos-López et al. [2016] and Montesinos-López et al. [2017] used a linear mixed model and Bayesian Poisson-lognormal method to predict multiple traits in multiple environments, explicitly considering $G \times E$ in their analysis. Lopez-Cruz et al. proposed a mixed model in which they explicitly took into account the $G \times E$. Cuevas et al. [2016] proposed a Gaussian kernel regression method incorporating a $G \times E$ mixed model and a single-environment model for prediction.

Several studies have applied machine learning techniques for crop yield prediction, including decision trees, multivariate regression, association rule mining, random forest, and artificial neural networks. Machine learning models treat the response variable (crop yield) as an implicit function of input variables such as genotype and environmental components Khaki and Wang [2019]. For instance, He et al. [2006] applied artificial neural networks to analyze the functional relationship between wheat yield and input variables, including nitrogen, organic matter, and water content. Guo and Xue [2014, 2012] used feed-forward neural networks and recurrent neural networks for crop yield prediction and evaluated their performance. Green et al. Green et al. [2007] related crop yield to topographic input variables such as elevation, curvature, and slope using neural networks and

multiple linear regression. Liu et al. Liu et al. [2001] trained a fully-connected neural network to approximate the nonlinear yield function based on input variables including soil, weather, and management. Drummond et al. Drummond et al. [2003] applied linear regression, projection pursuit regression, and neural networks to relate soil properties and grain yield, finding neural networks outperformed the other two methods. Romero et al. [2013] used decision tree (J48) and association rule mining (a priori) to predict durum wheat yield, finding the a priori method obtained overall the best performance Frawley et al. [1992]. Marko et al. [2016] proposed weighted histograms regression for yield prediction and compared their method with other regression algorithms.

Transformer models are a class of deep learning models that excel at processing sequential data by leveraging self-attention mechanisms to learn complex interactions among features in the data. They work by encoding input sequences and generating output sequences through attention-based mechanisms, enabling efficient learning of complex patterns in data without relying on recurrent connections. Recently, transformer based model have been used for crop yield prediction. For example, Onoufriou et al. [2023] proposed premonition network which is multi-timeline, time sequence ingesting approach based on transformer models towards processing the past, the present, and premonitions of the future for strawberry tabletop yield forecasting. Liu et al. [2022] proposed a transformer-based model, to predict rice yield by integrating time-series satellite data, environmental variables, and rice yield records from 2001 to 2016. They showed transformer models had better performance than four other machine learning and deep learning models for end-of-season prediction. Bi et al. [2023] utilized vision transformer-based approach for soybean yield prediction using early-stage images and seed information. Lin et al. [2023] developed a novel multi-modal spatial-temporal vision transformer model for predicting crop yields at the county level across the United States, by considering the effects of short-term meteorological variations during the growing season and the long-term climate change on crops. Krishnan et al. [2024] utilized transformer models for sugarcane yield prediction.

In this paper, we introduce novel Transformer-Enhanced Neural Networks models tailored for crop yield prediction, adept at efficiently considering diverse datasets including trait data,

metadata, soil data, weather data, genotype data, and APSIM data (environmental covariate (EC) data), thereby surpassing existing methodologies in prediction accuracy. The data under consideration encompasses various types of sequential information, including temporal patterns in weather data, spatial and temporal correlations within APSIM data, and genetic associations among adjacent genetic markers. Our proposed methodology adopts a modular approach by utilizing separate TransformerEncoder models tailored to each input data category. This entails dedicated transformers designed specifically for weather, EC phenological period-soil layer, EC-phenological period, and genotype data. Such a modular design enables our model to adapt flexibly to the distinct characteristics and patterns inherent in each data category. Each individual transformer model within our methodology is configured with two layers, attention heads, and a hidden size of 128.

Extrapolation is a crucial technique in predictive modeling, especially in the context of crop yield forecasting. It involves extending or projecting existing data points beyond the range of observed values to measure the model's capabilities and robustness to predict beyond the range of data it has seen during the training. In our study, with training data comprising different data types such as genotype, location, and year information, we used extrapolation to assess how well our proposed model performs when faced with new records including different years, hybrids, and locations that were not part of the training set. To this end, we preserved the records from the original test datasets, incorporating data from 2021 for temporal extrapolation. We then excluded records containing unique hybrids and field locations present in the test datasets for the first variable combinations involving genotype information (VC1, VC2, VC3, and VC4) from the corresponding training datasets. To ensure that the extrapolation analysis remained free from variables indicating specific years, hybrid names, and location names, we removed certain columns from both the training and test datasets. Finally, we evaluated the model's performance on these hold-out test datasets to get insights into its ability to generalize and make accurate predictions for unseen scenarios. This approach not only tests the robustness of our model but also provides valuable information for improving its predictive capabilities. In the literature, there are several

studies that used extrapolation to measure their model robustness for crop yield predictions. For example, Khaki et al. [2020] examined the power of their CNN-RNN model in generalizing the prediction to the locations that have never been trained on. They randomly excluded locations from the training data (1980–2017) and trained the CNN-RNN model on the remaining locations. Then, they tested the model on the excluded locations for the 2018 yield prediction.

This study is driven by three primary objectives. Firstly, it introduces Transformer-Enhanced Neural Networks as a solution for managing sequential data. Secondly, it delves into the examination of how different combinations of variables affect prediction errors, utilizing test data from the year 2021. Lastly, we extend our analysis to incorporate temporal, genomic, and geographic extrapolations. This is accomplished by training the Transformer-Enhanced Neural Networks models using newly curated training datasets tailored specifically for extrapolation analysis. Subsequently, we evaluate these models on the same test datasets used previously, with certain variables such as years, hybrid names, and location names excluded.

The subsequent sections show the structure of our paper: Section 4.3 explains the utilized data sources, Section 4.4 presents our innovative approach, Section 4.5 showcases the attained results, and Section 4.6 delves into additional analyses and insights derived from our proposed methodology. Finally, in Section 4.7, we wrap up the paper by discussing the contributions made by this study.

4.3 Data

In this study, we employed data sourced from the Genomes to Fields (G2F) Initiative Lima et al. [2023] spanning the years 2014-2021. The dataset encompassed phenotypic and genotypic information for 4,683 distinct hybrids assessed at 45 locations across the United States between 2014 and 2021. Additionally, comprehensive data on soil, weather, environmental covariates (EC), and metadata were collected for all environments, defined as the combination of year and location. The dataset comprised six sets of information: trait data, metadata, soil data, weather data, genotype data, and environmental covariate data (APSIM data). Each of the files related to trait data, metadata, soil data, weather data, and environmental covariate data included an "Env"

column, which served as a key for their integration. The "Hybrid" column served as a key for linking genotype data with the trait data and subsequently with other datasets. The "Env" column, short for Environment, was defined as the combination of the evaluation location and the corresponding year. All datasets were gathered starting from the year 2014, with the exception of soil data, which has been collected since 2015.

The trait data encompassed both plot and yield information for 4,683 distinct hybrids across 217 unique environments. Simultaneously, the metadata provided fundamental information about the locations and years associated with these 217 distinct environments. The soil dataset comprised essential soil data corresponding to various locations and years. It included 29 soil variables, the soil sample collection depth, as well as pertinent details such as the soil laboratory IDs responsible for sample analysis, the date of sample receipt at the laboratory, the date of sample processing reporting, and any comments provided by collaborators. Sixteen weather variables at a daily time scale were provided from the NASA Power website (https://power.larc.nasa.gov/) for locations spanning nine years (from 2014 to 2022). Simultaneously, the genotype dataset featured 4,928 distinct genotypes and 437,214 variant sites. The environmental covariate dataset encompassed 765 variables derived through an unpublished APSIM crop model developed by a team led by Aguate, Fernando; de Leon, Natalia; de los Campos, Gustavo; Holland, James; Kaeppler, Shawn; Lima, Dayane; Lopez-Cruz, Marco; Tan, Ruijuan; Thompson, Addie; and Washburn, Jacob. This simplified model uniformly applied 200 kg/ha of NO3 fertilization across all locations and utilized planting densities specified in the associated files. Phenological periods were estimated based on averages from the training data and were not specific to any particular hybrid. The names of environmental covariates were assigned by combining an EC-type, a phenological period, and a soil layer.

Among the 765 environmental covariates, 540 variables were associated with the combination of 6 EC-types, 9 phenological stages, and 10 soil layers. Additionally, 81 variables were linked to the combination of 1 EC-type (Flow), 9 phenological periods, and 9 soil layers, while 144 variables corresponded to the combination of 16 EC-types and 9 phenological stages. Phenological periods

consisted of 9 stages, including pGerEme (Germination to emergence), pEmeEnJ (Emergence to end of juvenile), pEnJFlo (End of juvenile to floral initiation), pFloFla (Floral initiation to flag leaf), pFlaFlw (Flag leaf to flowering), pFlwStG (Flowering to start of grain fill), pStGEnG (Start of grain fill to end of grain fill), pEnGMat (End of grain fill to maturity), and pMatHar (Maturity to harvest/ripe). Soil layers were delineated as 1 through 10 (1 being the top layer), with each layer representing 20cm in the soil column, resulting in a total depth of 2 meters. For instance, the column "SDR_pGerEme_1" contained environmental covariates based on the Water Supply-Demand Ratio (SDR) during the phenological period from Germination to emergence (pGerEme) within the topmost soil layer (soil layer 1).

The objective of the Genomes to Fields (G2F) Genotype by Environment Prediction Competition G2F Prediction Competition, 2022 was to forecast the performance of corn hybrids in the 2022 G2F trials using the existing G2F dataset. However, the ground truth response variables for 2022 were not disclosed post-competition. In this study, we employed the data spanning 2015-2020 (inclusive of all six sets of information) or 2014-2020 (comprising five sets of information, excluding soil data) as the training datasets, while the 2021 data served as the test dataset. All test samples represented unique environments (Envs), defined as combinations of evaluation location and the corresponding year, ensuring no overlap with the training data.

Table 4.1 presents a comprehensive summary of the datasets provided by the G2F prediction competition. These sets encompass trait data, metadata, soil data, weather data, genotype data, and environmental covariate data.

Table 4.1: Overview of the datasets provided by the G2F prediction competition, encompassing trait data, metadata, soil data, weather data, genotype data, and environmental covariate (EC) data.

Set of Information	**# of Unique Environments**	**# of Unique Hybrids**	**# of Variant Sites**	**Year**
Trait Data	217	4,683	N/A	2014-2021
Meta Data	217	N/A	N/A	2014-2021
Soil Data	141	N/A	N/A	2015-2021
Weather Data	212	N/A	N/A	2014-2021
Genotype Data	N/A	4,928	437,214	N/A
EC Data (APSIM Data)	165	N/A	N/A	2014-2021

We integrated six distinct datasets encompassing trait data, metadata, soil data, weather data, genotype data, and environmental covariate data. The pre-2021 data formed our training set, while the 2021 data served as our test set. The combined datasets provided comprehensive information for 28 locations in the training data and 13 locations in the test data. Fig 4.1 illustrates the geographic distribution of performance records across these cities in the United States. Each dot on the map corresponds to a location, with its size indicating the number of records, and its color reflecting the number of unique fields in each respective location.

4.4 Materials and methods

Given our primary goals of assessing a hybrid transformer-fully connected (FC) neural networks framework's performance and examining the impact of various combinations of soil, APSIM, and genotype variables, along with factors like weather, phenotypic, and metadata, this section is divided into two parts. The first outlines how we create various combinations of variables and delves into the specifics of the data we used, the associated pre-processing tasks, and the methods employed for variable creation. The second part elaborates on the prediction frameworks employed for different combinations of variables, covering models' inputs, details of selected predictive models, and the evaluation metrics used for result comparisons.

4.4.1 Data Preparation

4.4.1.1 Data Preprocessing

Geotype data A hybrid crop results from the crossbreeding of two inbred parents. The 437,214 genotype variables were encoded with 0, 1 values for the inbred parents. These inbred parents were crossed to generate genotype variables for the 4,928 hybrids in the genotype dataset, which were then merged with trait data using a shared hybrid column. In the crossing process, we assigned 0 if both parents were 0, 1 if both were 1, and a random number between 0 and 1 if one of the inbred parents was 1. Notably, we randomly selected 340 genotype variables from the total of 437,214 during the crossbreeding of inbred parents. Following the merger of genotype data and trait

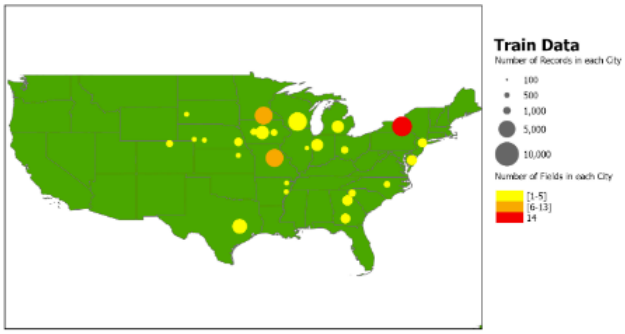
data using the common "hybrid" column, 260 hybrids present in the trait dataset were excluded from the genotype dataset. These exclusions are attributed to the unavailability of these hybrids in the genotype data, stemming from either their classification as commercial hybrids or the failure of their genotype data to meet quality control filters.

Weather data We undertook downsampling of the daily weather data by calculating the average and consolidating the feature values on a weekly basis for each of 16 weather variables. Recognizing the excessive detail and granularity inherent in daily data, which could impede knowledge discovery, we opted for a more manageable weekly information approach. This transformation resulted in a substantial reduction in the dimensionality of the weather data, maintaining a 365:52 ratio, thereby significantly decreasing the model complexity. The preprocessing step of downsampling daily weather data to a weekly level aligns with established practices in yield prediction studies, providing a more tractable and informative dataset for analysis Khaki and Wang [2019], Srivastava et al. [2022], Khalilzadeh et al. [2023a].

Environmental covariate (EC) data (APSIM data) As outlined in section 4.3, the APSIM dataset comprised 621 variables (540+81), intricately linked to the combination of EC-types, phenological stages, and soil layers, with an additional 144 variables corresponding to the combination of EC-types and phenological stages. To capture the spatial and temporal dependencies inherent in these variables, we decided to organize them into two distinct groups of APSIM variables within the dataset. These groups consist of EC variables providing information for 9 phenological stages across 10 or 9 soil layers, and EC variables exclusively associated with the 9 phenological stages. Termed as EC-phenological period-soil layer (ECPS) and EC-phenological period (ECP), respectively, this division enabled their integration into separate transformer models. This approach ensures a focused and specialized treatment for each set of variables, enhancing the model's ability to comprehend and represent the intricacies within the APSIM dataset.

Merging Process During the merging process, various combinations of available datasets were merged, including weather data, APSIM data, genotype data, trait data, metadata, and soil data, resulting in the creation of 8 base merged datasets. Each iteration of merging involved the deliberate exclusion of specific datasets, which yielded the advantage of increasing the number of records at the expense of losing certain variables within each dataset. For example, excluding soil data increased the number of observations because the soil data was from 2015, while other datasets were from 2014. This exclusion ensured that we retained the records from 2014 without compromising the dataset's integrity. Similarly, omitting genotype data preserved records for 260 hybrids, as genotype information was provided for only 4423 hybrids out of the total 4683 in the trait dataset. Additionally, removing APSIM data augmented the number of records while relinquishing APSIM variables. This increase occurred because the APSIM dataset contained 165 unique environments. After removing APSIM data, we retained records with uncommon unique environments found in other datasets. Consequently, the combined dataset, which includes trait data and metadata (217 unique environments) as well as weather data (212 unique environments), contributed to the augmentation of the number of records. Our approach involved a column-wise integration strategy, leveraging a shared column 'hybrid' to merge genotype data (comprising 340 variables) with trait data. Subsequently, the 'Env' column served as the key for merging this consolidated dataset with other datasets including preprocessed weather data, featuring weekly values of 16 weather variables for 52 weeks, APSIM data, Metadata, and soil data. The merging process was executed in several stages: In the first merging, all available datasets including genotype, trait data, metadata, soil data, weather data, and APSIM data were merged. In the second merging, soil data was excluded from the first merging. In the third merging, APSIM data was excluded from merging, and all other datasets were merged. In the fourth merging, both APSIM and soil data were excluded from the merging process. In the fifth merging, genotype data was excluded, and the remaining datasets were merged. In the sixth merging, both genotype and soil data were excluded. In the seventh merging, genotype and APSIM data were excluded. Finally,

in the eighth merging, genotype data, APSIM data, and soil data were excluded, and the remaining datasets including weather, trait data, and metadata were merged.



(a)

(b)

Figure 4.1: The distribution of performance records across 28 locations in the train data (a), and 13 locations in the test data (b) within the U.S. states. The size of each dot represents the number of records, and the color of the dot corresponds to the number of unique fields in each respective location.

Missing Values Following the creation of the combined datasets as described, we conducted data cleaning by excluding variables with over 30 percent missing values. Additionally, records with missing values in the target variable, Yield_Mg_ha, representing grain yield in Mg per ha at 15.5% grain moisture, focusing on plot areas without alleys (Mg/ha), were removed.

Imputing Missing Values For variables exhibiting less than 30 percent missing values, we employed imputation techniques to address the missing values. Specifically, for categorical variables, we used the mode, and for numerical variables, we used the median as the imputation method.

Excluding Unnecessary Variables To streamline data and simplify model complexity, we strategically omitted redundant variables, including the latitudes and longitudes of field corners and weather stations, soil sample processing dates, weather station placement dates, and redundant hybrid information such as parent names and original hybrid names.

Handling Categorical Variables with One-Hot Encoding We opted to treat certain numerical variables as categorical and, when including them alongside other categorical variables, applied one-hot encoding. In this encoding scheme, each unique value of every categorical variable is represented as a new binary feature in a separate column. The numerical variables considered as categorical include: Replicate (Large-scale field block), Block (Smaller-scale field block nested within Replicate), Plot (Designation of the individual experimental unit), Range (Designation of the field range of the plot, organized perpendicular to corn rows), Pass (Designation of the field pass of the plot, organized parallel to corn rows; a combination of range and pass forms a coordinate grid system describing the location of each plot within the field), Year (Year of evaluation), WDRF Buffer pH (Woodruff method for measuring total soil acidity), and Texture No (Particle size analysis with mineral components smaller than 2mm).

Handling Numerical Variables with Z-Score Normalization Given the diverse range of values and varying scales present in numerical variables, it is crucial to mitigate potential biases originating from individual features. To address this concern, we implemented the z-score

normalization technique (Equation 4.1) to standardize all numerical values. This process transforms all numerical variables to adhere to a standard normal distribution, thereby preventing unintentional biases in the results. Beyond bias mitigation, standardizing numerical variable values also enhances the numerical robustness of the models and expedites training speed.

$$V_{i,j} = \frac{v_{i,j} - \bar{v}_j}{\sigma_j} \tag{4.1}$$

Where $V_{i,j}$ represents the standardized value of the i-th observation of the j-th numerical variable (j ranges from 1 to K, where K is the total number of numerical variables. For our specific datasets, K is as follows: 765 for APSIM variables, 340 for genotype variables, 832 for weather variables, 21 for soil data, and 10 for trait and metadata variables.), $v_{i,j}$ is the original value of the i-th observation of the j-th numerical variable, $\bar{v}_j$ is the mean of the j-th numerical variable, and σ_j is the standard deviation of the j-th numerical variable. The formula rescales each variable to have a mean of 0 and a standard deviation of 1.

4.4.1.2 Variable Combinations

In this study, our focus extended to assessing the combined effects of different variables. To achieve this, we crafted eight unique combinations of datasets. To create these eight unique combinations of datasets we used the eight datasets which was created using merging different datasets as it was explained in the paragraph Merging Process in Subsection 4.4.1.1. These initial merging processes gave rise to eight variable combinations as follows:

1. First Variable Combination: Utilizing data from the first merging, weather, genotype, and APSIM (including ECPS and ECP) datasets were separated for inputs to transformer models. Simultaneously, input_other was created, including trait data, metadata, and soil data, intended for input to the fully-connected neural networks.

2. Second Variable Combination: Similar to the first combination, using data from the second merging, weather, genotype, and APSIM datasets were separated for transformer models, and input_other included trait data and metadata for the fully-connected neural networks.

3. Third Variable Combination: Derived from the third merging, in which APSIM data was excluded, resulting in separate weather and genotype datasets. Input_other, containing trait, metadata, and soil data, was retained.

4. Fourth Variable Combination: Replicating the third, but using data from the forth merging, input_other included trait data and metadata.

5. Fifth Variable Combination: Created from the fifth merging, separate datasets for weather and APSIM were generated for input to transformer models. Input_other retained trait, metadata, and soil data.

6. Sixth Variable Combination: Similar to the fifth but using data from the sixth merging, input_other included trait data and metadata.

7. Seventh Variable Combination: Derived from the seventh merging, in which genotype and APSIM datasets were excluded, resulting in a separate weather dataset. Input_other retained trait, metadata, and soil data.

8. Eighth Variable Combination: Replicating the seventh, but using data from the eighth merging, input_other included trait data and metadata.

4.4.2 Model development

In constructing our proposed transformer-fully connected (Transformer-FC) models, we strategically leveraged different combinations of datasets to harness the temporal dependencies in weather data, spatial and temporal correlations in APSIM data, and genetic linkages among adjacent genetic markers. Recognizing the capacity of transformer models to capture such dependencies, we designed four distinct sets of two-layer transformer models for weather features (W-transformer), APSIM features including EC-phenological period-soil layer(APS-transformer), and EC-phenological period (AP-transformer), and genotype features (G-transformer). The input_other data, which includes trait data, metadata, and the option to include or exclude soil

data, was used as input for our fully connected layer. Consequently, we explored various combinations of variables and model architectures, resulting in eight distinct combinations.

1. Model 1: Four separate transformer models for Weather, APSIM (ECPS and ECP), and Genotype data.

   One fully connected model for `input_other` data, including soil data.

2. Model 2: Four separate transformer models for Weather, APSIM (ECPS and ECP), and Genotype data.

   One fully connected model for `input_other` data, excluding soil data.

3. Model 3: Two separate transformer models for Weather and Genotype data (excluding APSIM data).

   One fully connected model for `input_other` data, including soil data.

4. Model 4: Two separate transformer models for Weather and Genotype data (excluding APSIM data).

   One fully connected model for `input_other` data, excluding soil data.

5. Model 5: Three separate transformer models for Weather, and APSIM (ECPS and ECP) (excluding Genotype data).

   One fully connected model for `input_other` data, including soil data.

6. Model 6: Three separate transformer models for Weather, and APSIM (ECPS and ECP) (excluding Genotype data).

   One fully connected model for `input_other` data, excluding soil data.

7. Model 7: One transformer model for only weather data (excluding genotype and APSIM data).

   One fully connected model for `input_other` data, including soil data.

8. Model 8: One transformer model for only weather data (excluding Genotype and APSIM data).

One fully connected model for `input_other` data, excluding soil data.

4.4.2.1 Proposed Transformer-Enhanced Neural Networks

Our proposed method utilizes transformer models [Vaswani et al., 2023] to adeptly handle sequential data. The sequential data under consideration includes temporal dependencies in weather data, spatial and temporal correlations in APSIM data, and genetic linkages among adjacent genetic markers. Transformers are a type of deep learning model that excels in processing sequential data. At the core of transformers is the self-attention mechanism, which allows the model to weigh different parts of the input sequence differently when making predictions. Multi-head attention in transformers involves using multiple attention mechanisms in parallel, enabling the model to capture various aspects and dependencies within the input data simultaneously, enhancing its ability to understand complex relationships and patterns.

Our proposed methodology adopts a modular approach, employing distinct TransformerEncoder models for each input data category. This includes dedicated transformers for weather, EC-phenological period-soil layer, EC-phenological period, and genotype data. This modular design allows our model to adapt to the unique characteristics and patterns present in each data category. Each individual transformer model in our methodology is configured with two layers, 8 attention heads, and a hidden size of 128. Additionally, we integrate a dropout layer with a rate of 10%. Dropout is a regularization technique used during training, where randomly selected neurons are ignored. This helps prevent overfitting and improves the model's generalization performance. By including dropout layers, we enhance the robustness of our models during training, mitigating overfitting and promoting improved generalization performance across diverse agricultural datasets. The concept of attention heads refers to the parallel attention mechanisms within the transformer model. Each attention head focuses on different aspects of the input sequence, allowing the model to attend to various patterns simultaneously. The use of 8 attention heads enhances the model's ability to capture nuanced relationships and dependencies within the data. The hidden size is a crucial parameter determining the dimensionality of the model's internal

representation. In our methodology, the hidden size is set to 128, providing a balance between model complexity and computational efficiency. This carefully chosen dimensionality allows the model to encode intricate features and relationships in the data while maintaining computational efficiency during training and inference. The choice of two layers ensures the model's ability to capture both local and global dependencies in the sequential input data.

Input_other data is fed into a fully−connected neural network with three layers. The architecture of this network involves three consecutive linear transformation layers with Rectified Linear Unit (ReLU) activation functions. The number of hidden units in these layers is set to 128. This design choice aims to capture intricate patterns and relationships within the input data. Notably, the inclusion of the ReLU activation function introduces non-linearity, enhancing the model's capacity to learn complex representations. The fully-connected neural network structure ensures the transformation of input features into a meaningful representation for subsequent stages of the model.

The high−level features from the transformer models are then concatenated with the output of the fully−connected neural network. The combined features are then processed through one additional FC layer. The linear layer takes the concatenated output features from the individual models and fully connected layer, refining them into the conclusive prediction for corn yield.

Figures 4.2, 4.3, 4.8, and 4.5 show the modeling architecture of our proposed method considering various combinations of inputs for the transformer models.

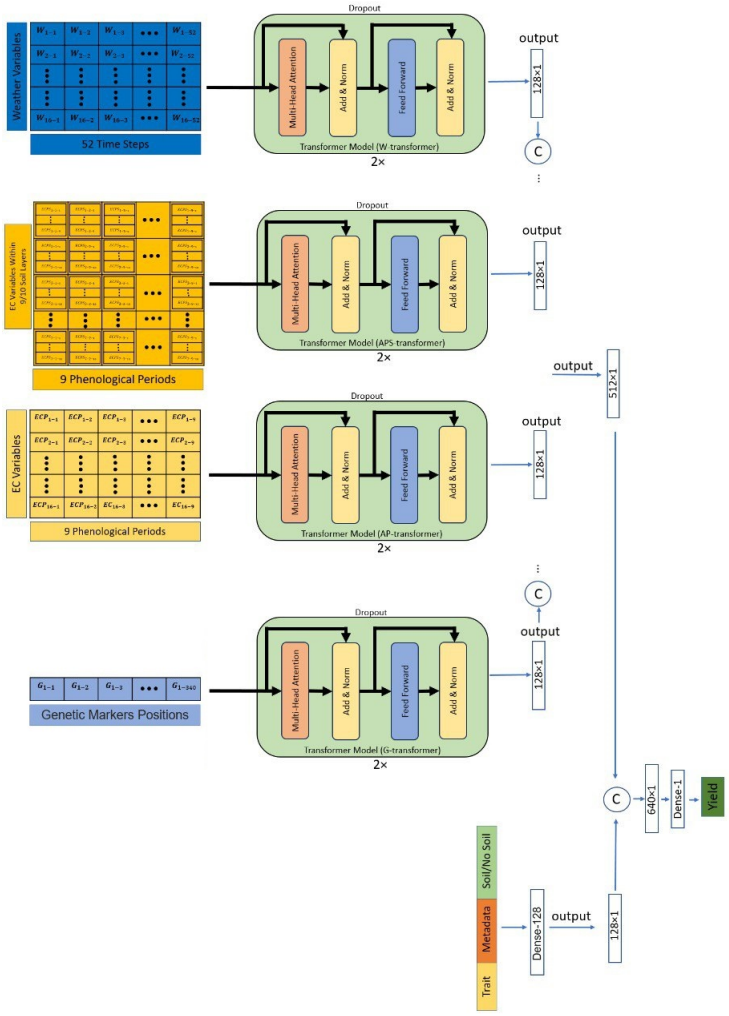


Figure 4.2: Models 1 and 2: Four separate transformer models for Weather, APSIM (ECPS and ECP), and Genotype data. One fully connected model for input_other data, including/excluding soil data. The fully connected layers are labeled as Dense, and matrix concatenations are represented by Ⓒ. Yield represents the final corn yield prediction made by the model.

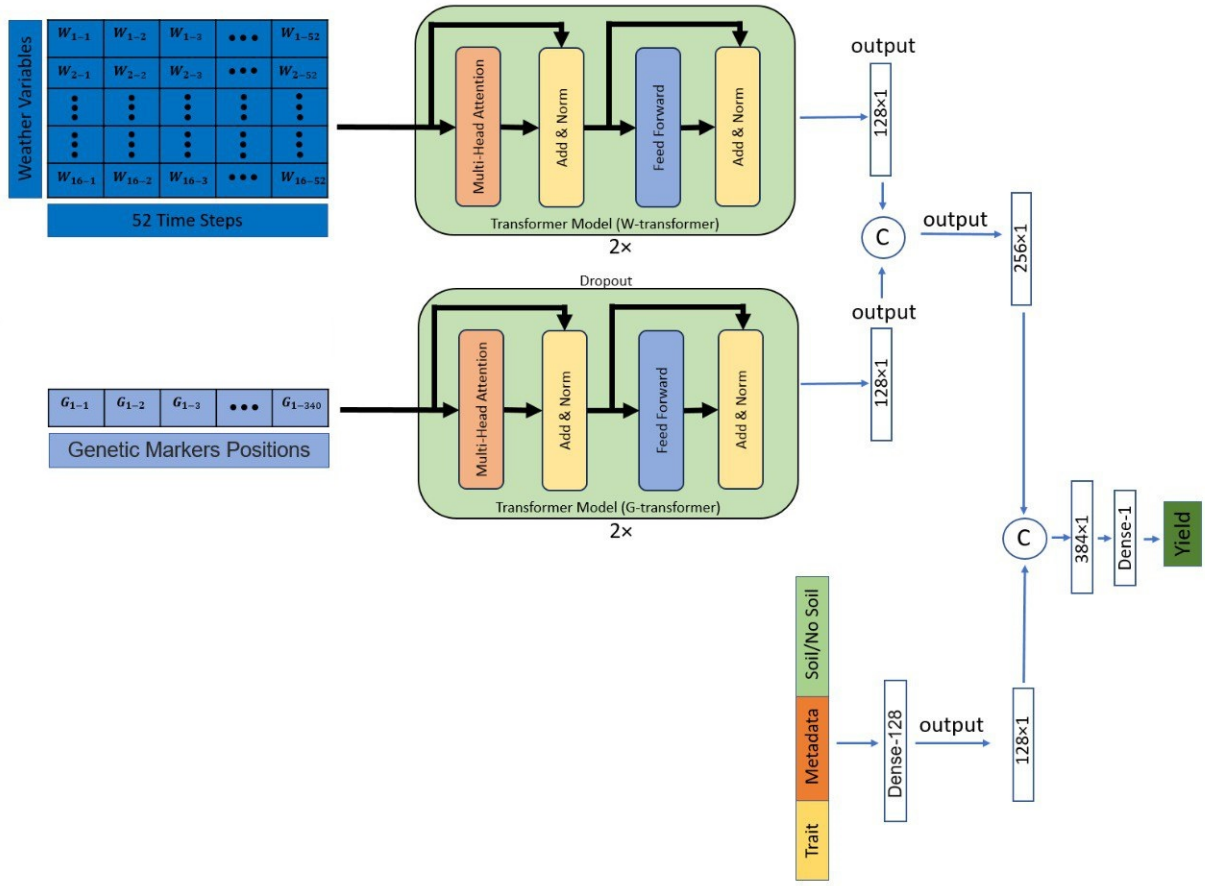


Figure 4.3: Models 3 and 4: Two separate transformer models for Weather and Genotype data (excluding APSIM data). One fully connected model for input_other data, including/excluding soil data. The fully connected layers are labeled as Dense, and matrix concatenations are represented by Ⓒ. Yield represents the final corn yield prediction made by the model.

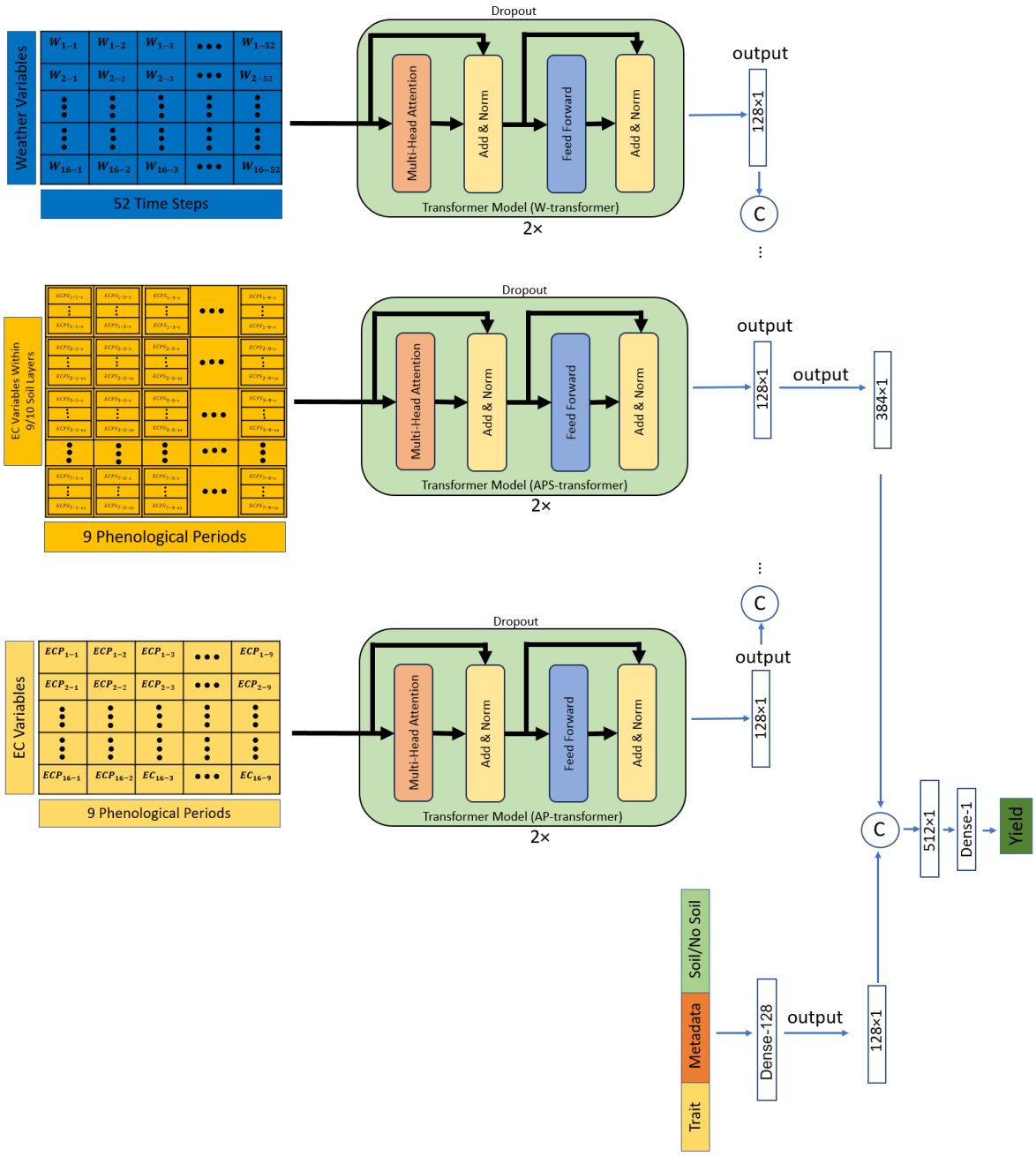


Figure 4.4: Models 5 and 6: Three separate transformer models for Weather, and APSIM (ECPS and ECP) (excluding Genotype data). One fully connected model for input_other data, including/excluding soil data. The fully connected layers are labeled as Dense, and matrix concatenations are represented by Ⓒ. Yield represents the final corn yield prediction made by the model.

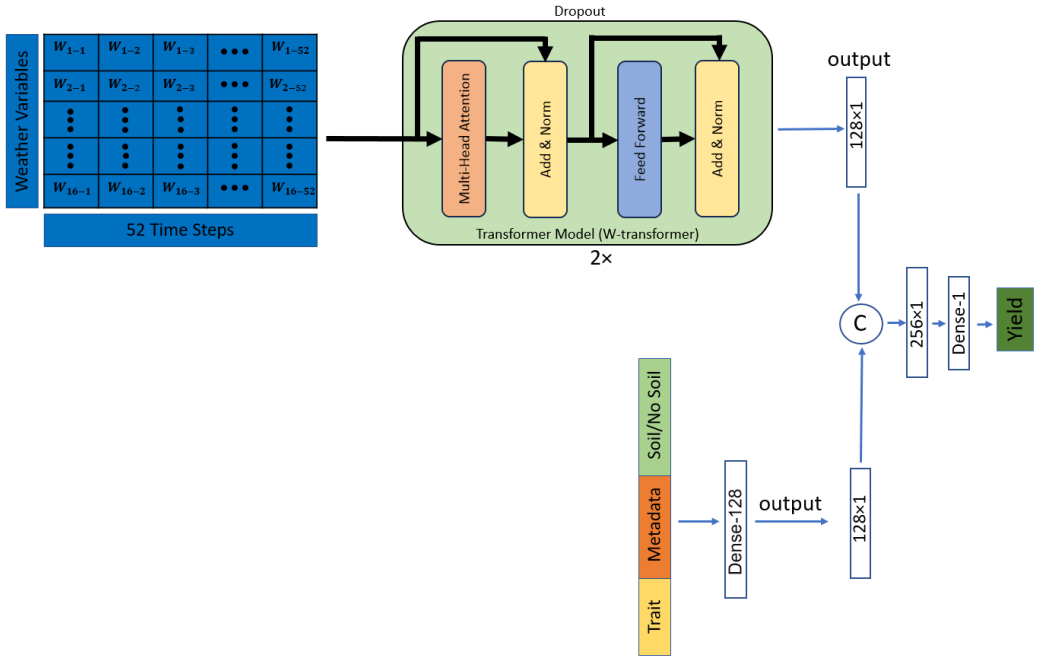


Figure 4.5: Models 7 and 8: One transformer model for only weather data (excluding Genotype and APSIM data). One fully connected model for input_other data, including/excluding soil data. The fully connected layers are labeled as Dense, and matrix concatenations are represented by Ⓒ. Yield represents the final corn yield prediction made by the model.

4.4.2.2 Design of Experiments

In this study, we utilized data from the years 2014 to 2020, encompassing diverse combinations of all six sets of information, or from 2015 to 2020, including various combinations of five sets of information (excluding soil data), as the training datasets. The data from the year 2021 was designated as the test dataset.

We employed the following hyperparameters to train our Transformer-Enhanced Neural Networks models. Each of the four transformer components (W-transformer, APS-transformer, AP-transformer, and G-transformer) comprises two layers, 8 attention heads, a hidden size of 128, and incorporates a dropout layer set at a rate of 10%. Our designed Fully Connected Neural Network consists of three fully connected (dense) layers, each featuring a linear transformation.

The input features connect to a hidden layer, where the hidden size is configured as 128; subsequently, the hidden layers are interconnected. Following each linear transformation, a Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity to the model. We initialized the weights using the PyTorch default weight initialization method, specifically the Xavier initialization method [Glorot and Bengio, 2010]. Stochastic Gradient Descent (SGD) was employed with a mini-batch size of 128. The Adam optimizer [Kingma and Ba, 2014] was utilized with a learning rate set to 1e-4. The model underwent training for a total of 30 epochs. The Rectified Linear Unit (ReLU) activation function was applied in the fully connected (FC) layer, while the output layer featured a linear activation function. Implementation of the proposed model was carried out in Python using the PyTorch library [Paszke et al., 2019]. Computations were performed on a Google Colab instance equipped with an NVIDIA Tesla T4 GPU.

In order to compare transformer models with one-dimensional convolutional neural networks (CNNs), we employed a high-performing CNN-deep neural network (DNN) model proposed by Khalilzadeh et al. [2023b]. The CNN architecture underwent a slight modification, and instead of using valid padding for all layers, we employed the same padding for APSIM (ECPS and ECP) layers and the initial layer in the CNN model for weather and genotype variables. Instead of using transformer models for Weather, APSIM (ECPS and ECP), and Genotype data, as outlined in Models 1 to 8 in the subsection 4.4.2, we opted for separate CNN models. These models integrate CNNs and fully-connected (FC) neural networks. Utilizing a diverse set of input features including weather variables, APSIM ECPS variables, APSIM ECP variables, and genotype variables, these models comprise up to four distinct CNN components: W-CNN, ECPS-CNN, ECP-CNN, and G-CNN. These components adeptly capture spatial and temporal dependencies, employing one-dimensional convolution operations to unravel intricate linear and nonlinear effects within the input data. These models then concatenate their corresponding outputs to capture high-level features. The input_others, encompassing trait data, metadata, and both inclusive and exclusive representations of soil data, are directed into a single-layer fully-connected neural network. This dense layer takes the input input_others, performs a linear transformation using weights initialized

with Glorot normal initializer, applies L2 regularization to the weights, and includes bias terms. The layer has 2048 units and does not apply an activation function, implying it is a purely linear transformation. Such layers, common in neural network architectures, are pivotal for learning complex mappings between input and output, facilitating the model's ability to capture complex relationships within the data. The concatenated high-level features from the CNN models are combined with the output of the fully-connected neural network for input_others data. These combined features are processed through two additional FC layers before yielding the final corn yield prediction. The first layer transforms the combined features using a fully connected (dense) layer with 3200 units, applying the ReLU activation function. The Glorot normal initializer initializes the weights, and L2 regularization is employed to mitigate overfitting. The second layer takes the output from the previous layer and performs a linear transformation, producing a single output unit. The deliberate omission of an activation function implies a linear activation, rendering it particularly well-suited for regression tasks. To prevent overfitting, three dropout layers with dropout ratios of 0.5, 0.7, and 0.2 respectively are added after the CNN layers, at the end of the fully-connected layer for input_others data, and at the final layer of the model. The architecture of the CNN-DNN models for eight variable combinations are described in Figures 4.6, 4.7, 4.8, and 4.9. The details of the CNN networks including W-CNN, ECPS-CNN, ECP-CNN, and G-CNN are presented in Tables 4.2, 4.3, 4.4, and 4.5, respectively. The models were trained using the Adam optimizer with a scheduled learning rate of 0.0004, decaying exponentially with a rate of 0.96 every 2500 steps. Training comprised 100,000 iterations with a batch size of 48. The activation function chosen for all networks was Rectified Linear Unit (ReLU), with the exception of the fully-connected layer associated with input_other data, which operated without an activation function.

To further evaluate our proposed Transformer-Enhanced Neural Networks models, we used six machine learning models including Least Absolute Shrinkage and Selection Operator (LASSO) regression, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF), K-Nearest Neighbors (KNN), and Regression Tree (RT). LASSO regression is a linear regression technique that incorporates regularization to prevent overfitting and

perform feature selection. In LASSO, the objective is to minimize the sum of squared errors, subject to the constraint that the sum of the absolute values of the regression coefficients is less than a predefined constant. This constraint encourages sparsity in the model, effectively setting some coefficients to exactly zero [James et al., 2013]. The strength of regularization is controlled by a hyperparameter, often denoted as alpha. LASSO is particularly useful when dealing with high-dimensional datasets, as it helps identify and prioritize the most influential features while mitigating multicollinearity [Tibshirani, 1996]. XGBoost is a powerful and versatile machine learning algorithm belonging to the family of gradient boosting methods. It builds a predictive model by combining the outputs of multiple weak learners, typically decision trees, to create a robust and accurate ensemble model. XGBoost employs a unique regularization term in its objective function, enhancing its ability to handle complex relationships and outliers. The algorithm iteratively adds trees to the model, with each tree addressing the errors of the previous ones [Chen and Guestrin, 2016]. LightGBM is a gradient boosting framework that shares similarities with XGBoost but introduces optimizations to enhance training speed and efficiency, making it well-suited for large-scale datasets. One key innovation is the implementation of a histogram-based learning approach, where data is binned to facilitate faster and more memory-efficient training. LightGBM also supports distributed computing, making it suitable for parallel and distributed environments [Ke et al., 2017]. RF is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree is built independently, utilizing a random subset of the training data and a random subset of features at each split. This randomness helps decorrelate the trees and reduce overfitting [Cutler et al., 2007]. KNN is a simple yet effective machine learning algorithm used for both classification and regression tasks. The fundamental idea behind KNN is to predict the target value of an unseen data point based on the values of its 'k' nearest neighbors in the feature space [Cover and Hart, 1967]. The determination of proximity is typically based on a distance metric, and in this study we opted for the Euclidean distance. The Euclidean distance between two points, $\mathbf{X}$ and $\mathbf{Y}$, in an n-dimensional

space is computed using the following formula:

$$\text{Distance}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.2}$$

Here, x_i and y_i represent the corresponding feature values of points $\mathbf{X}$ and $\mathbf{Y}$ in the n-dimensional space. The Euclidean distance essentially measures the straight-line distance between these points. RT is a regression-focused model built on decision tree principles. It partitions the feature space into distinct regions and assigns a constant value, which, in our study, corresponds to the mean of the target values within each region. The tree's structure evolves through recursive data splits based on feature thresholds, optimizing a specific criterion—in our case, mean squared error. RTs are interpretable, handle non-linearity well, and can capture complex relationships.

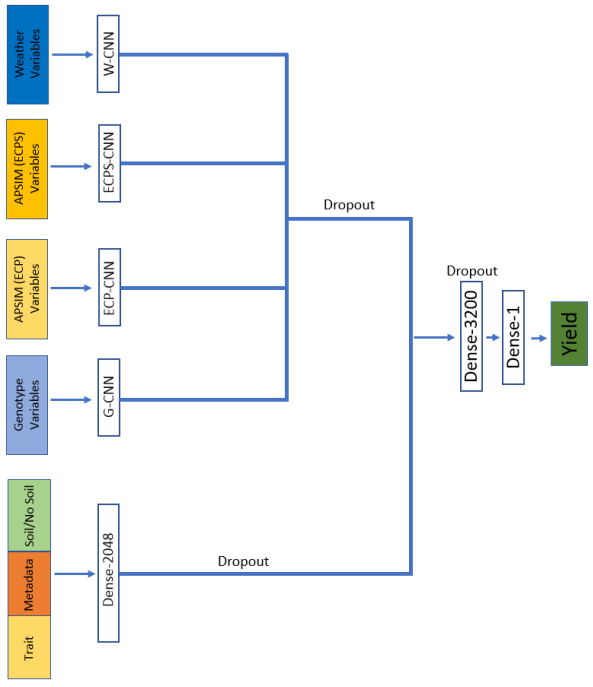


Figure 4.6: The CNN-DNN architecture for the First and Second Variable Combinations. Four separate CNN models for weather, APSIM (ECPS and ECP), and genotype data. One fully connected model for input_other data, including/excluding soil data. Yield represents the final corn yield prediction made by the model.

To train the mentioned ML models we used Scikit-learn library in Python. Scikit-learn, often abbreviated as sklearn, is a comprehensive machine learning library for Python. It provides simple and efficient tools for data analysis and modeling, including various algorithms for classification, regression, clustering, dimensionality reduction, and more. Scikit-learn is built on NumPy, SciPy, and Matplotlib, making it seamlessly integrate with the broader Python data science ecosystem. Scikit-learn supports a wide range of machine learning tasks and includes functionalities for data preprocessing, model evaluation, and hyperparameter tuning [Pedregosa et al., 2011].

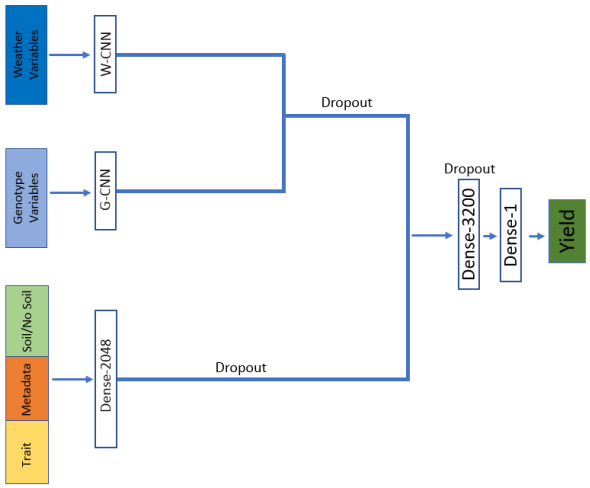


Figure 4.7: The CNN-DNN architecture for the Third and Fourth Variable Combinations. Two separate CNN models for weather and genotype data (excluding APSIM data). One fully connected model for input_other data, including/excluding soil data. Yield represents the final corn yield prediction made by the model.

For LASSO regression, the LassoCV class in scikit-learn was utilized, leveraging the default 5-fold cross-validation to determine the optimal regularization strength during the training process [scikit-learn development team, 2023c]. The alpha parameter in LassoCV governs the regularization strength, and the algorithm systematically explores a range of alpha values, selecting the one that maximizes the default scoring metric coefficient of determination R^2 [scikit-learn development team, 2023c]. During the cross-validation, we examined 100 default alpha values, and the determination

of alpha values is based on the epsilon length of the path, with a default value of 0.001, indicating that $\frac{\text{alpha\_min}}{\text{alpha\_max}} = 1 \times 10^{-3}$ [scikit-learn development team, 2023c]. The optimal alpha parameters for each variable combination are presented in Table 4.6. In our analysis, we utilized scikit-learn-compatible APIs, employing xgboost.XGBRegressor and lightgbm.LGBMRegressor to train XGBoost and LightGBM models, respectively, in Python. RF, KNN, and RT models were trained using RandomForestRegressor, KNeighborsRegressor, and DecisionTreeRegressor respectively from the scikit-learn library. Default hyperparameters were employed for XGBoost, LightGBM, RF, KNN, and RT as outlined in their respective documentation sources [Documentation, 2022], [Team, 2024], [scikit-learn development team, 2023d], [scikit-learn development team, 2023b], and [scikit-learn development team, 2023a].

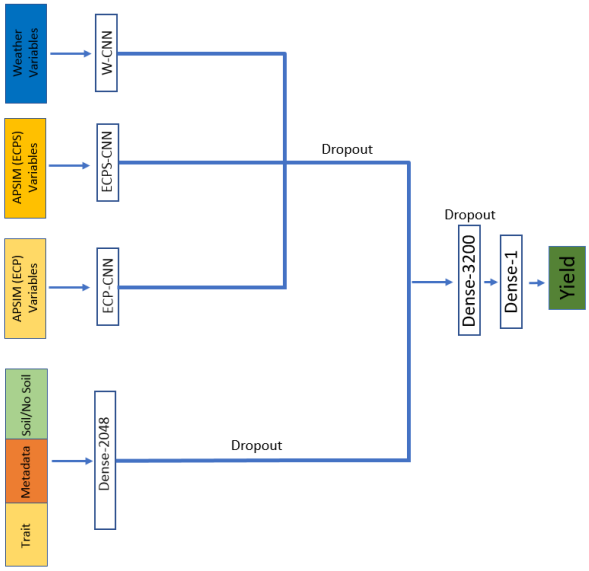


Figure 4.8: The CNN-DNN architecture for the Fifth and Sixth Variable Combinations. Three separate CNN models for weather, and APSIM (ECPS and ECP) (excluding genotype data). One fully connected model for input_other data, including/excluding soil data. Yield represents the final corn yield prediction made by the model.

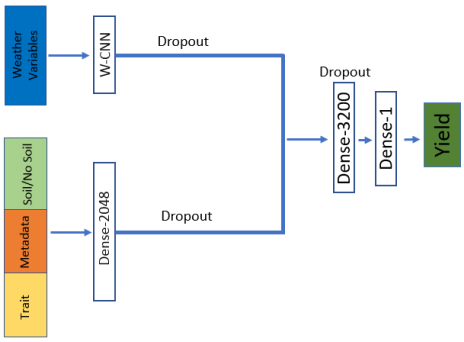


Figure 4.9: The CNN-DNN architecture for the Seventh and Eighth Variable Combinations. One transformer model for only weather data (excluding genotype and APSIM data). One fully connected model for input_other data, including/excluding soil data. Yield represents the final corn yield prediction made by the model.

Table 4.2: CNN in the W-CNN component of the models. FS, NF, S, and P stand for filter size, number of filter, stride, and padding, respectively.

Input Size	52×16			
Layer names	**FS**	**NF**	**S**	**P**
Conv 1	3	24	3	same
Conv 2	3	48	2	valid
Conv 3	3	64	1	valid
Conv 4	3	128	2	valid
Global Avg. Pooling	-	-	-	-
Output Size	1×128			

Table 4.3: CNN in the ECPS-CNN component of the models. FS, NF, S, and P stand for filter size, number of filter, stride, and padding, respectively.

Input Size	9×69			
Layer names	**FS**	**NF**	**S**	**P**
Conv 1	3	24	3	same
Conv 2	3	48	2	same
Conv 3	3	64	1	same
Conv 4	3	128	2	same
Output Size	1×128			

Table 4.4: CNN in the ECP-CNN component of the models. FS, NF, S, and P stand for filter size, number of filter, stride, and padding, respectively.

Input Size	9×16			
Layer names	**FS**	**NF**	**S**	**P**
Conv 1	3	24	3	same
Conv 2	3	48	2	same
Conv 3	3	64	1	same
Conv 4	3	128	2	same
Output Size	1×128			

Table 4.5: CNN in the G-CNN component of the models. FS, NF, S, and P stand for filter size, number of filter, stride, and padding, respectively.

Input Size	340 × 1			
Layer names	**FS**	**NF**	**S**	**P**
Conv 1	3	24	3	same
Conv 2	3	48	2	valid
Conv 3	3	64	1	valid
Conv 4	3	128	2	valid
Global Avg. Pooling	-	-	-	-
Output Size	1 × 128			

Table 4.6: Alpha values for LASSO regression across variable combinations.

Combination 1	Combination 2	Combination 3	Combination 4	Combination 5	Combination 6	Combination 7	Combination 8
0.223	0.292	0.302	0.075	0.277	0.478	0.303	0.327

4.4.2.3 Performance Metrics

In this study, the performance of the prediction models was rigorously assessed using two widely acknowledged metrics: Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2). These metrics provide comprehensive insights into the accuracy and explanatory power of the models, respectively.

Root Mean Square Error (RMSE) RMSE is a prevalent metric in regression analysis, quantifying the average magnitude of prediction errors. It is calculated as the square root of the mean squared differences between the predicted (y^{pred}) and actual (y^{true}) values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{\text{true}} - y_i^{\text{pred}})^2} \tag{4.3}$$

Coefficient of Determination (R^2) The R^2 metric, also known as the coefficient of determination, assesses the proportion of the variance in the dependent variable that is explained by the model. In other words, it quantifies the goodness of fit of the model to the data. R^2 values range from 0 to 1. A value of 1 indicates a perfect fit, meaning the model perfectly predicts the dependent variable. A value of 0 indicates that the model does not explain any variability in the dependent variable.

The formula for R^2 is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{\text{true}} - y_i^{\text{pred}})^2}{\sum_{i=1}^{n}(y_i^{\text{true}} - \bar{y}^{\text{true}})^2} \tag{4.4}$$

where:

- y_i^{true} is the true value of the dependent variable for observation i,
- y_i^{pred} is the predicted value of the dependent variable for observation i,
- $\bar{y}^{\text{true}}$ is the mean of the true values,
- n is the number of observations.

4.5 Results

In this section, we investigate the impact of various combinations of variables on prediction errors using the test data for the 2021 year. This analysis is conducted across the proposed Transformer-Enhanced models, 1D convolutional neural networks, and six machine learning models.

4.5.1 Performance of Models for Different Variable Combinations

In this subsection, we systematically explore the impact of different variable combinations on the prediction performance of our proposed Transformer-Enhanced Neural Networks models, 1D convolutional neural networks (CNN-DNN), and six ML models including LASSO regression, XGBoost, LightGBM, RF, KNN, and RT. Eight distinct variable combinations have been considered, as detailed in Subsection 4.4.1.2. Each combination derived from distinct merging of soil daset with weather, genotype, APSIM (including ECPS and ECP), trait data, and metadata.

These combinations aim to comprehensively investigate the model's response to diverse input configurations, including the inclusion and exclusion of genotype data, APSIM (ECPS and ECP) data, and soil data (as described in Subsection 4.4.1.1). These eight variable combinations (VC) encompass the following configurations:

VC1: Derived from the first merging, VC1 includes weather, APSIM (ECPS and ECP), genotype data, trait data, metadata, and soil data.

VC2: Derived from the second merging, VC2 includes weather, APSIM (ECPS and ECP), genotype data, trait data and metadata.

VC3: Derived from the third merging, VC3 includes weather, genotype data, trait data, metadata, and soil data.

VC4: Derived from the fourth merging, VC4 includes weather, genotype data, trait data, and metadata.

VC5: Derived from the fifth merging, VC5 includes weather, APSIM (ECPS and ECP) data, trait data, metadata, and soil data.

VC6: Derived from the sixth merging, VC6 includes weather, APSIM (ECPS and ECP) data, trait data, and metadata.

VC7: Derived from the seventh merging, VC7 includes weather data, trait data, metadata, and soil data.

VC8: Derived from the eighth merging, VC8 includes weather data, trait data, and metadata.

Tables 4.7, and 4.8 present summary statistics across various variable combinations for test data and train data respectively.

Table 4.7: Summary statistics of test data for the variable combinations (VC) 1 to 8. The unit of the corn yield is Mg per ha at 15.5% grain moisture.

Summary statistics	**VC 1**	**VC 2**	**VC 3**	**VC 4**	**VC 5**	**VC 6**	**VC 7**	**VC 8**
Mean yield	10.084	10.100	10.023	10.040	10.083	10.101	10.027	10.044
Standard deviation of yield	3.025	2.927	2.836	2.779	3.029	2.932	2.840	2.783
25th percentile of yield	8.598	8.672	8.495	8.558	8.595	8.671	8.501	8.561
Median yield	10.519	10.477	10.320	10.316	10.522	10.482	10.326	10.319
75th percentile of yield	12.131	12.045	11.909	11.871	12.136	12.051	11.918	11.879
Minimum yield	0.584	0.584	0.584	0.584	0.584	0.584	0.584	0.584
Maximum yield	18.766	18.766	18.766	18.766	18.766	18.766	18.766	18.766
Number of observations	12,243	13,890	17,814	19,461	12,319	13,972	17,924	19,577

Table 4.8: Summary statistics of train data for the variable combinations (VC) 1 to 8. The unit of the corn yield is Mg per ha at 15.5% grain moisture.

Summary statistics	**VC 1**	**VC 2**	**VC 3**	**VC 4**	**VC 5**	**VC 6**	**VC 7**	**VC 8**
Mean yield	9.203	9.185	9.447	9.412	9.218	9.198	9.465	9.425
Standard deviation of yield	2.972	2.913	3.074	2.997	2.981	2.921	3.083	3.005
25th percentile of yield	7.276	7.258	7.431	7.409	7.283	7.262	7.442	7.413
Median yield	9.223	9.231	9.471	9.470	9.241	9.244	9.492	9.484
75th percentile of yield	11.267	11.209	11.585	11.502	11.288	11.229	11.607	11.519
Minimum yield	0.548	0.500	0.548	0.500	0.548	0.500	0.548	0.500
Maximum yield	22.798	22.798	22.798	23.268	22.798	22.798	22.798	23.268
Number of observations	55,564	78,990	69,244	104,056	57,041	80,854	71,197	106,579

Table 4.9 provides a comparative analysis of the performance of the proposed Transformer-Enhanced models, CNN-DNN model, and six ML models on the test and train datasets. The evaluation is conducted across eight variable combinations (VC1:VC8), considering

both RMSE and R^2 for corn yield prediction. The outcomes indicate a significant superiority of the proposed Transformer-Enhanced models over the other six ML models across various variable combinations. This is more evident in Fig. 4.10. Particularly, when considering weather, and genotype data while excluding APSIM (ECPS and ECP), and soil data from the merging process (VC4), the Transformer-Enhanced model exhibits the most impressive performance. This affirms the model's ability to effectively handle sequential data. Despite the omission of the APSIM and soil datasets during the merging process, resulting in the loss of ECPS, ECP, and soil variables, this strategic choice enables the incorporation of more observations. It allowed us to include more observations starting from 2014 when we excluded soil data. Additionally, by excluding APSIM data, we ended up with more records because we incorporated more unique environments. This trade-off helped us increase the overall dataset size despite losing certain variables. Among all baseline models, the LASSO model demonstrates superior performance across all performance metrics. Notably, when excluding soil data and APSIM datasets (including ECPS and ECP) in VC4, the LASSO model outperforms other variable combinations. Beyond VC4, VC8, which further excludes soil, APSIM (including ECPS and ECP), and genotype datasets, attains the highest performance across all performance measures. A negative R^2 signifies that ML models, including Random Forest, and Regression Tree across all variable combinations, as well as XGBoost across all variable combinations except VC2, VC4, and VC7, and LightGBM across all variable combinations except VC3, VC4, VC7, and VC8, and KNN across all variable combinations except VC2, and VC4 perform worse than a naive mean-based model. This suggests that the predictions from these models are notably poor, making it more favorable to use the mean of the dependent variable as a predictor. This observation indicates that these ML models, for the mentioned variable combinations, fail to capture any discernible patterns in the data or produce predictions systematically inferior to the mean. The findings also imply the existence of an overfitting problem in the prediction results of certain baseline models such as Random Forest and Regression Tree.

Table 4.9: Performance comparison of different models across variable combinations (VC).

Model	Test															
	VC 1		VC 2		VC 3		VC 4		VC 5		VC 6		VC 7		VC 8	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Transformer-Enhanced	2.433	0.304	2.301	0.302	2.272	0.300	**2.202**	0.303	2.348	0.349	2.298	0.311	2.248	0.313	2.251	0.275
CNN-DNN	3.222	-0.135	2.929	-0.001	2.774	0.043	2.585	0.135	3.385	-0.248	2.891	0.028	2.874	-0.024	2.786	-0.002
LASSO	2.992	0.022	2.814	0.076	2.655	0.124	2.513	0.182	2.966	0.041	2.836	0.064	2.673	0.114	2.557	0.156
XGBoost	3.139	-0.077	2.887	0.027	2.917	-0.058	2.742	0.026	3.296	-0.184	3.017	-0.059	2.760	0.055	2.829	-0.033
LightGBM	3.157	-0.089	3.015	-0.061	2.713	0.085	2.589	0.131	3.174	-0.098	3.023	-0.063	2.735	0.072	2.594	0.132
Random Forest	3.283	-0.178	3.350	-0.310	3.010	-0.127	2.932	-0.113	3.294	-0.182	3.257	-0.234	2.981	-0.101	2.872	-0.065
KNN	3.108	-0.055	2.825	0.069	2.877	-0.029	2.717	0.044	3.249	-0.149	3.178	-0.175	3.349	-0.390	2.959	-0.130
Regression Tree	4.023	-0.768	4.315	-1.172	4.105	-1.095	4.354	-1.455	3.861	-0.625	3.928	-0.795	4.102	-1.086	4.084	-1.153
Model	Train															
	VC 1		VC 2		VC 3		VC 4		VC 5		VC 6		VC 7		VC 8	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Transformer-Enhanced	1.878	0.598	1.700	0.657	1.551	0.743	1.544	0.733	1.847	0.614	1.646	0.680	1.724	0.685	1.678	0.686
CNN-DNN	1.713	0.668	1.515	0.730	1.786	0.662	1.482	0.755	1.640	0.697	1.450	0.753	1.785	0.665	1.245	0.828
LASSO	2.352	0.374	2.546	0.236	2.613	0.277	2.327	0.397	2.454	0.322	2.691	0.151	2.647	0.263	2.687	0.201
XGBoost	1.505	0.743	1.479	0.742	1.494	0.764	1.512	0.745	1.569	0.723	1.551	0.718	1.581	0.737	1.582	0.723
LightGBM	1.592	0.713	1.581	0.705	1.599	0.729	1.597	0.716	1.648	0.694	1.630	0.688	1.652	0.713	1.650	0.698
Random Forest	0.597	0.959	0.575	0.961	0.588	0.963	0.566	0.964	0.618	0.957	0.596	0.958	0.614	0.960	0.591	0.961
KNN	1.532	0.734	1.484	0.740	1.544	0.747	1.526	0.741	1.488	0.751	1.432	0.759	1.479	0.769	1.424	0.775
Regression Tree	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

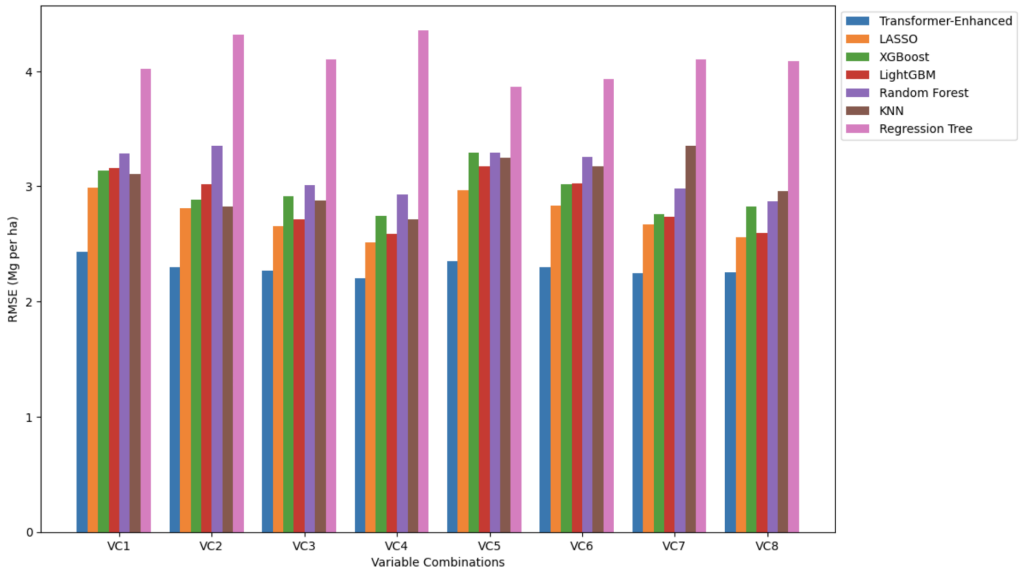


Figure 4.10: Evaluation of RMSE performance across Transformer-Enhanced and baseline ML models for each variable combination (VC) using test data.

Analyzing RMSE performance across diverse variable combinations for each ML model using test data (Fig. 4.11), our findings consistently demonstrate that the effectiveness of ML models, measured through RMSE, undergoes improvement with the gradual exclusion of datasets like soil, APSIM (including ECPS and ECP), and genotype datasets (VC7 and VC8). This advancement is attributed to the strategic omission of soil data during the merging process, facilitating an augmented number of observations while concurrently diminishing the variables. Moreover, the intricate nature of sequential data, encompassing APSIM (including ECPS and ECP) and genotype datasets, presents a challenge that standalone ML models may encounter difficulties in addressing. Additionally, it is noteworthy that not all variables within these datasets contribute valuable information to ML yield predictions. Moreover, the results indicate that excluding the APSIM and soil datasets while incorporating genotype variables has had a significant positive impact. Specifically, all ML models, except for RT and RF, demonstrate improved performance in terms of RMSE when VC4 is used. This suggests that the strategic decision to exclude APSIM and soil

datasets while including genotype variables has led to enhanced model performance across various ML algorithms.

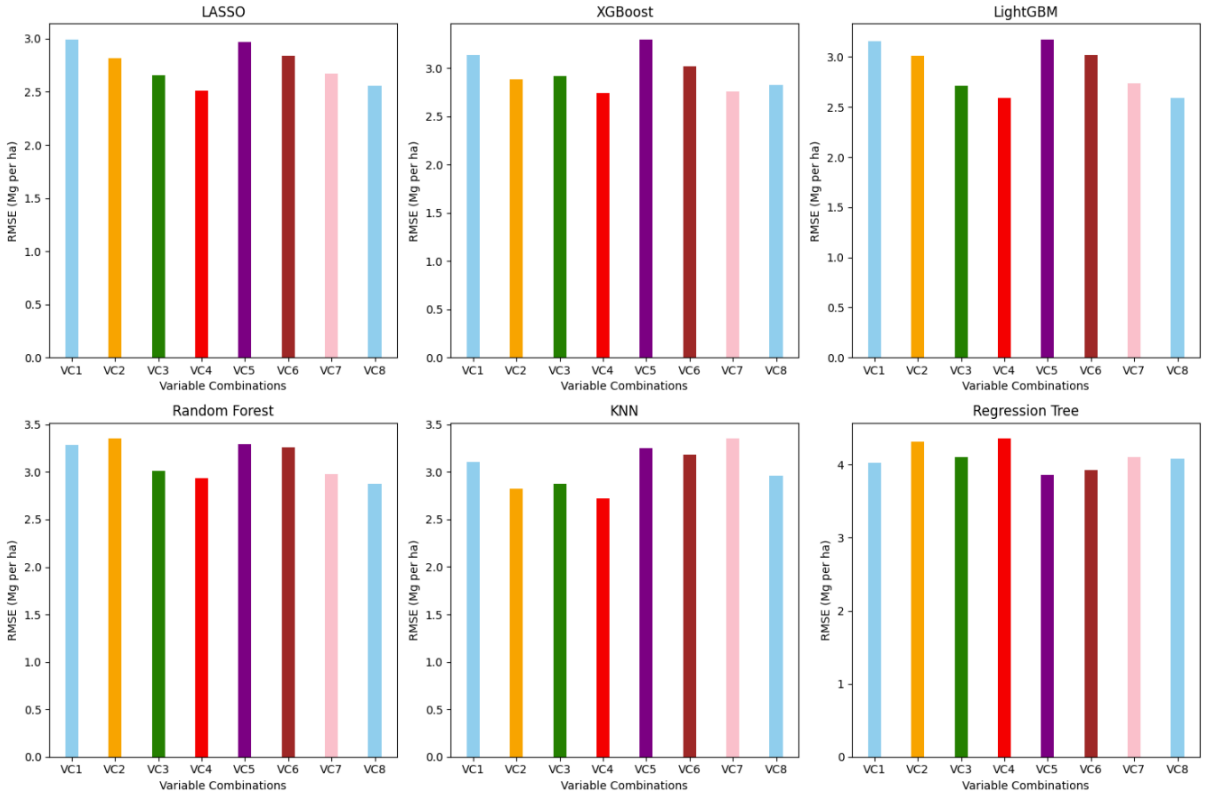


Figure 4.11: RMSE performance across various variable combinations (VC) for each ML model using test data.

As it was discussed in 4.4.2, 8 Tranformer-Enhanced models were developed to deal with sequential data in each of 8 variable combinations. For the CNN-DNN models instead of using transformer layers for each of weather variables, APSIM ECPS variables, APSIM ECP variables, and genotype variables, these variables are separately taken as input to the convolutional neural network part of the model. Table 4.9 shows that our proposed model outperforms the CNN-DNN model with respect to both RMSE and R^2 in all variable combinations. A negative R^2 signifies that CNN-DNN model's predictions across VC1, VC2, VC5, VC7, and VC8 are performing worse than a

simple model that predicts the mean value of the dependent variable. In other words, the models are not providing any meaningful explanatory power, and their performance is even poorer than a basic average-based prediction.

Looking at the comparison of RMSE performance of Transformer-Enhanced and CNN-DNN Models across diverse variable combinations (Fig.4.12), it is evident that the proposed Transformer-Enhanced models excel in handling sequential data. Notably, for variable combination 5 (VC5), encompassing all datasets except genotype data, and subsequently for VC1, comprising all datasets, the model exhibits the most substantial decrease in RMSE, 44% and 32%, respectively. In both models, VC4, which includes partial sequential variables such as weather variables and genotype variables, and excludes soil data from the input_other data, demonstrates optimal performance with a larger dataset. This suggests that weather variables and genotype variables significantly influence the predictions made by both Transformer-Enhanced and CNN-DNN models. Additionally, it underscores the importance of these variables in providing valuable information for deep learning models used in yield prediction. For the Transformer-Enhanced model, following VC4, VC7, and VC3 exhibited the highest performance. These variable combinations include soil data and contain partial sequential data, with VC7 consisting of weather variables and VC3 incorporating genotype and weather variables. In the case of the CNN-DNN model, VC3 demonstrated the highest performance after VC4. VC3 shares the same partial sequential data as VC4, including genotype and weather variables, while also incorporating soil data in input_other data. Following VC3, VC8 achieved the highest performance. VC8 excludes soil data, genotype variables, and APSIM (ECPS and ECP) variables. The performance of both models suggests that incorporating APSIM data into the variable combination has a detrimental effect. Specifically, the results show that variable combinations VC1, VC2, VC5, and VC6, which include APSIM data, exhibit the lowest performance among all configurations.

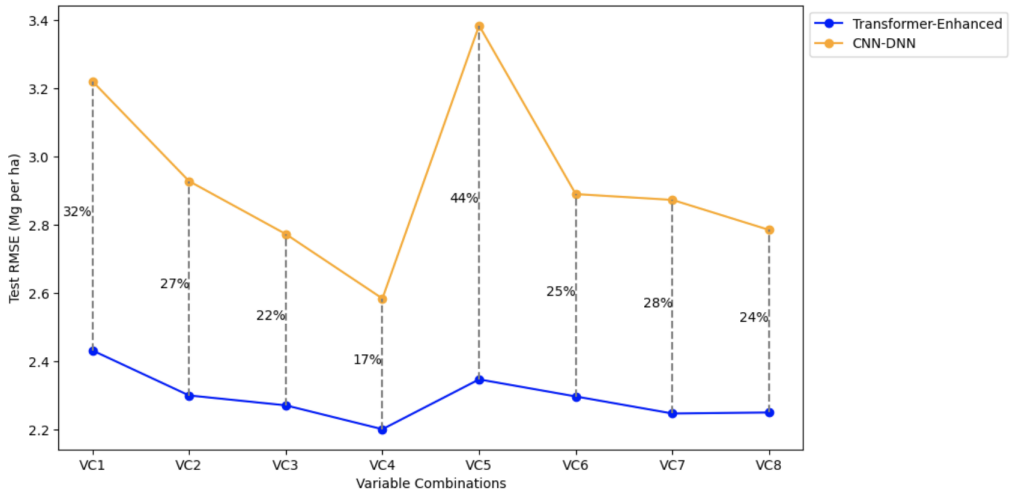


Figure 4.12: Comparison of RMSE performance of Transformer-Enhanced and CNN-DNN models across diverse variable combinations (VC) using test data.

4.6 Analysis

In this section, we extend our analysis to include temporal, genomic, and geographic extrapolations to assess the robustness of our proposed Transformer-Enhanced Neural Networks models across different variable combinations.

4.6.1 Extrapolation Analysis

In crop yield prediction, temporal, genomic, and geographic extrapolation refer to different aspects of extending predictive models beyond the scope of the data used to train them:

Temporal extrapolation: This involves predicting future crop yields based on historical data. Temporal extrapolation extends the predictive model to forecast yields for time periods beyond those covered by the training data.

Genomic extrapolation: This refers to predicting crop yields for genotypes or genetic variants that were not present in the training data. Genomic extrapolation involves applying predictive models trained on genotype-phenotype data to new genotypes or genetic variants that were not

included in the original dataset. This can be particularly useful in plant breeding programs, where genomic information is used to develop new crop varieties.

Geographic extrapolation: This involves predicting crop yields for geographic locations or regions that were not represented in the training data. Geographic extrapolation extends the predictive model to make predictions for areas beyond those covered by the training data.

From the eight variable combinations considered in our study, we focused on the four variable combinations (VC1, VC2, VC3, and VC4) that included genotype information and performed extrapolation analysis by incorporating genomic and geographic factors in addition to temporal information. In the previous part, we created the test data for all variable combinations based solely on temporal extrapolation, predicting future crop yields (2021) using historical data from 2014-2020 or 2015-2020. For the extrapolation part, we augmented the analysis with genomic and geographic factors, resulting in the creation of new train datasets. We retained the records from the original test datasets (including data from 2021 for temporal extrapolation) and excluded records including unique hybrids and field locations present in the test datasets for each variable combination (VC1, VC2, VC3, and VC4) from the corresponding train datasets. To ensure that the extrapolation analysis does not include variables indicating specific years, hybrid names, and location names, we removed certain columns from both the training and test datasets. These columns include `Experiment`, `Env`, `Field_Location`, `City`, `Farm`, `Year`, `LabID`, `Hybrid`, `Weather_Station_Serial_Number (Last four digits, e.g.  m2700s#####)`, and `Experiment_Code`, depending on the variable combination. Some other variables indicating year, hybrid name, and location names were removed due to having more than 30% missing values in the new train datasets. The newly created train datasets consist of subsets of records and variables from the original train datasets, hence they are labeled as VC′1, VC′2, VC′3, and VC′4. Summary statistics across these newly created train datasets for extrapolation analysis for VC′1, VC′2, VC′3, and VC′4 are presented in Table 4.10. Notably, the records in the test datasets remain unchanged; however, for this section, certain variables were removed to facilitate extrapolation analysis.

Consequently, the summary statistics remain consistent with those of the original test datasets for VC1 to VC4, as provided in Table 4.7.

Table 4.10: Summary statistics across the created train datasets for extrapolation analysis for VC′1, VC′2, VC′3, and VC′4. The unit of the corn yield is Mg per ha at 15.5% grain moisture.

Summary statistics	**VC′1**	**VC′2**	**VC′3**	**VC′4**
Mean yield	8.104	8.405	7.951	8.347
Standard deviation of yield	2.783	2.869	2.858	2.871
25th percentile of yield	6.367	6.444	6.105	6.356
Median yield	8.212	8.459	8.062	8.432
75th percentile of yield	9.991	10.480	9.904	10.430
Minimum yield	0.548	0.532	0.678	0.532
Maximum yield	22.798	22.798	20.945	23.268
Number of observations	15,411	18,089	15,193	22,198

The Transformer-Enhanced models were tested on the same test datasets (but excluding variables indicating specific years, hybrid names, and location names) as before where we considered only temporal extrapolation but trained using newly created train datasets tailored for extrapolation analysis. Detailed performance results of these models, derived from the extrapolation-oriented train datasets and tested on the modified test datasets, are summarized in Table 4.11.

Table 4.11: Performance comparison of extrapolation analysis using the proposed Transformer-Enhanced Neural Networks models across VC′1 to VC′4.

	Test							
Model	VC′1		VC′2		VC′3		VC′4	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Transformer-Enhanced	2.849	0.015	2.567	0.099	**2.498**	0.150	2.501	0.113
	Train							
Model	VC′1		VC′2		VC′3		VC′4	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Transformer-Enhanced	1.660	0.641	1.625	0.676	1.866	0.571	1.807	0.602

In the previous section, we found that VC4 had the lowest RMSE when considering only temporal extrapolation. VC4 is similar to VC3 and includes weather, genotype data, trait data, metadata, but doesn't include soil data. However, in the current analysis where we considered temporal, genomic, and geographic factors, VC′3 emerged as the best performer as it is shown in Table 4.11. VC′3 includes weather, genotype, and soil data but excludes APSIM variables. This result is interesting because using VC′3 means we have fewer historical records due to including soil data. Despite this, it highlights the importance of using diverse data types for more accurate predictions in extrapolation analysis.

The results highlight the robustness and capability of our proposed Transformer-Enhanced models in handling sequential data. Our advanced Transformer-Enhanced models, evaluated on the same datasets (with certain details omitted) but trained differently by using curated datasets specifically designed for extrapolation analysis, demonstrated promising performance. Despite being trained on less detailed data, our models experienced only marginal increases in prediction errors (RMSE) across variable combinations considered for the extrapolation analysis, as illustrated in Fig

4.13. Specifically, the RMSE increased by only 17%, 12%, 10%, and 12% for VC1 and VC′1, VC2 and VC′2, VC3 and VC′3, and VC4 and VC′4, respectively.

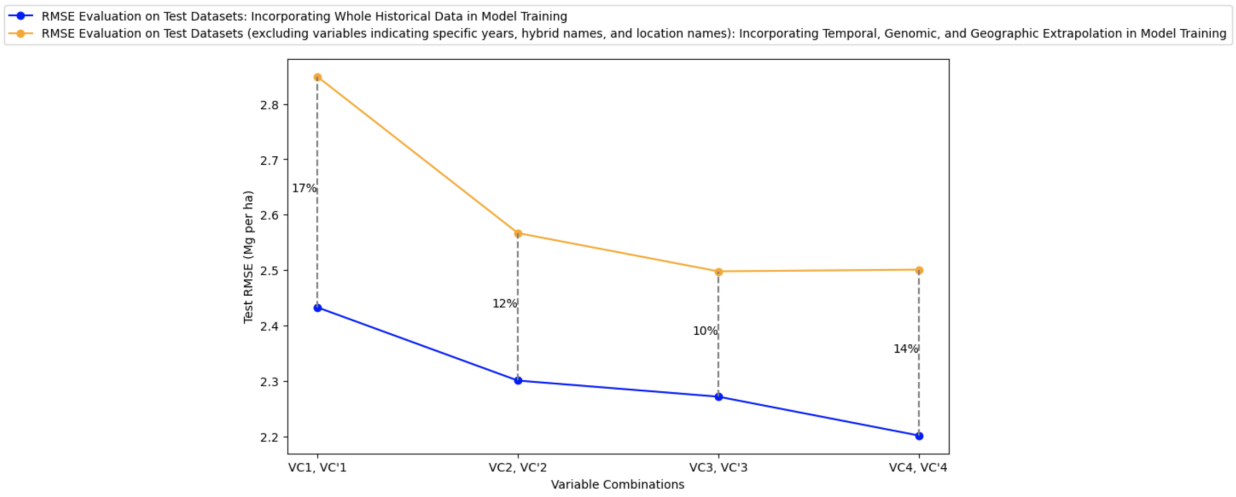


Figure 4.13: Comparison of RMSE performance for Transformer-Enhanced models: temporal, genomic, and geographic extrapolation vs. temporal extrapolation across variable combinations considered for the extrapolation analysis (VC1, VC′1, VC2, VC′2, VC3, VC′3, and VC4, VC′4) using original and trimmed test data.

4.7 Conclusion

In this study, we proposed to use a hybrid transformer-fully connected neural networks framework to adeptly handle sequential data for crop yield prediction. The sequential data under consideration included temporal dependencies in weather data, spatial and temporal correlations in APSIM data, and genetic linkages among adjacent genetic markers. Recognizing the inherent ability of transformer models to capture such intricate dependencies, we designed four distinct sets of two-layer transformer models for different data types: weather features (W-transformer), APSIM features incorporating EC-phenological period and soil layer information (APS-transformer), APSIM features focusing solely on EC-phenological period (AP-transformer), and genotype features

(G-transformer). Additionally, we incorporated input_other data, which encompassed trait information, metadata, and the option to include or exclude soil data, as inputs for the fully-connected layer part of the models, comprising three layers and 128 hidden units.

Additionally, we investigated the impact of various combinations of variables on prediction errors using test data for the year 2021. The study utilized a systematic approach to merge various datasets, including genotype, trait, metadata, soil, weather, and APSIM data, to evaluate the combined effects of different variables. This process resulted in eight unique combinations of datasets, each used to develop a specific model: Model 1: Utilized four separate transformer models for weather, APSIM (ECPS and ECP), and genotype data, alongside one fully connected model for additional data including soil. Model 2: Similar to Model 1 but excluded soil data from the inputs of the fully connected model. Model 3: Employed two separate transformer models for weather and genotype data (excluding APSIM), with a fully connected model for additional data including soil. Model 4: Similar to Model 3 but excluded soil data from the inputs of the fully connected model. Model 5: Included three separate transformer models for weather and APSIM (ECPS and ECP) data (excluding genotype), with a fully connected model for additional data including soil. Model 6: Similar to Model 5 but excluded soil data from the inputs of the fully connected model. Model 7: Utilized one transformer model for weather data only (excluding genotype and APSIM), alongside a fully connected model for additional data including soil. Model 8: Similar to Model 7 but excluded soil data from the inputs of the fully connected model. The results demonstrated a notable performance of the proposed Transformer-Enhanced models over six baseline ML models across diverse variable combinations. This underscored the model's effectiveness in handling sequential data. Particularly, the Transformer-Enhanced models showcased superior performance on VC4, which encompassed weather and genotype data while excluding APSIM (ECPS and ECP) and soil data. Our analysis of RMSE performance across various variable combinations for each ML model using test data consistently showed that the effectiveness of ML models, as measured by RMSE, improved as datasets such as soil, APSIM (including ECPS and ECP), and genotype datasets (VC7 and VC8) were gradually excluded. This improvement was attributed to the strategic omission of

soil data during the merging process, which allowed for an increased number of observations while reducing variables. Furthermore, the complex nature of sequential data, including APSIM (including ECPS and ECP) and genotype datasets, posed challenges that standalone ML models may struggle to address. Additionally, not all variables within these datasets contributed valuable information to ML yield predictions. The results also indicated a significant positive impact of excluding the APSIM and soil datasets while incorporating genotype variables (VC4), and all ML models, apart from RT and RF, demonstrated enhanced performance in terms of RMSE when VC4 was used. This indicated that the strategic choice to exclude APSIM and soil datasets while incorporating genotype variables has resulted in enhanced model performance across a variety of ML models.

We also conducted a comparison between our Transformer-Enhanced Neural Networks models and one-dimensional convolutional neural networks (CNNs). Instead of using transformer layers for sequential variables, we employed separate CNN models, including W-CNN, ECPS-CNN, ECP-CNN, and G-CNN, for weather, EC-phenological period and soil layer information, EC-phenological period, and genotype data. The results demonstrated that the proposed Transformer-Enhanced models excelled in effectively managing sequential data. Particularly noteworthy was the significant decrease in RMSE observed for VC5, which included all datasets except genotype data, and subsequently for VC1, comprising all datasets. In these cases, the proposed model exhibited the most substantial decreases in RMSE, by 44% and 32%, respectively.

Finally, the analysis was extended to include temporal, genomic, and geographic extrapolations to assess the robustness of the proposed Transformer-Enhanced models across different variable combinations. The results highlighted that our proposed Transformer-Enhanced models effectively generalize yield predictions to untested years, hybrids, and locations.

4.8 References

Bi, L., Wally, O., Hu, G., Tenuta, A. U., and Mueller, D. S. (2023). A transformer-based approach for early prediction of soybean yield using time-series images. *Frontiers in Plant Science*, 14:1173036.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cooper, M. and DeLacy, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88(5):561–572.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., Campos, G. d. l., Montesinos-López, O., and Burgueño, J. (2016). Genomic prediction of genotype× environment interaction kernel regression models. *The Plant Genome*, 9(3).

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

DeLacy, I., Basford, K., Cooper, M., Bull, J., McLaren, C., et al. (1996). Analysis of multi-environment trials–an historical perspective. *Plant adaptation and crop improvement*, 39124:39–124.

Denis, J. b. (1988). Two way analysis using covarites1. *Statistics*, 19(1):123–132.

Documentation, X. (2022). Xgboost parameters. Accessed: 01/07/2024.

Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., and Kitchen, N. R. (2003). Statistical and neural methods for site–specific yield prediction. *Transactions of the ASAE*, 46(1):5.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57.

G2F Prediction Competition, 2022. Genomes to Fields (G2F) Genotype by Environment Prediction Competition.

Gauch Jr, H. G. (2006). Statistical analysis of yield trials by ammi and gge. *Crop science*, 46(4):1488–1500.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Green, T. R., Salas, J. D., Martinez, A., and Erskine, R. H. (2007). Relating crop yield to topographic attributes using spatial analysis neural networks and regression. *Geoderma*, 139(1-2):23–37.

Guo, W. W. and Xue, H. (2012). An incorporative statistic and neural approach for crop yield modelling and forecasting. *Neural Computing and Applications*, 21(1):109–117.

Guo, W. W. and Xue, H. (2014). Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models. *Mathematical Problems in Engineering*, 2014.

Hammer, G., Kropff, M., Sinclair, T., and Porter, J. (2002). Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. *European Journal of Agronomy*, 18(1-2):15–31.

He, Y., Zhang, Y., Zhang, S., and Fang, H. (2006). Application of artificial neural network on relationship analysis between wheat yield and soil nutrients. In *2005. IEEE-EMBS 2005. 27th Annual International Conference of the Engineering in Medicine and Biology Society*, pages 4530–4533. IEEE.

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and applied genetics*, 127(2):463–480.

Hongyu, K., García-Peña, M., de Araújo, L. B., and dos Santos Dias, C. T. (2014). Statistical analysis of yield trials by ammi analysis of genotype× environment interaction. *Biometrical letters*, 51(2):89–102.

Horie, T., Yajima, M., and Nakagawa, H. (1992). Yield forecasting. *Agricultural Systems*, 40(1-3):211–236.

Jame, Y. and Cutforth, H. (1996). Crop growth models for decision support systems. *Canadian Journal of Plant Science*, 76(1):9–19.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Khaki, S. and Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621.

Khaki, S., Wang, L., and Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750.

Khalilzadeh, Z., Kashanian, M., Khaki, S., and Wang, L. (2023a). A hybrid deep learning-based approach for optimal genotype by environment selection. *arXiv preprint arXiv:2309.13021*.

Khalilzadeh, Z., Kashanian, M., Khaki, S., and Wang, L. (2023b). A hybrid deep learning-based approach for optimal genotype by environment selection.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishnan, V. G., Rao, B. S., Prasad, J. R., Pushpa, P., and Kumari, S. (2024). Sugarcane yield prediction using noa-based swin transformer model in iot smart agriculture. *Journal of Applied Biology and Biotechnology*, 12(2):239–247.

Lima, D. C., Washburn, J. D., Varela, J. I., Chen, Q., Gage, J. L., Romay, M. C., Holland, J., Ertl, D., Lopez-Cruz, M., Aguate, F. M., et al. (2023). Genomes to fields 2022 maize genotype by environment prediction competition. *BMC Research Notes*, 16(1):148.

Lin, F., Crawford, S., Guillot, K., Zhang, Y., Chen, Y., Yuan, X., Chen, L., Williams, S., Minvielle, R., Xiao, X., et al. (2023). Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5774–5784.

Liu, J., Goering, C., and Tian, L. (2001). A neural network for setting target corn yields. *Transactions of the ASAE*, 44(3):705.

Liu, Y., Wang, S., Chen, J., Chen, B., Wang, X., Hao, D., and Sun, L. (2022). Rice yield prediction and model interpretation based on satellite and climatic indicators using a transformer method. *Remote Sensing*, 14(19):5045.

Marko, O., Brdar, S., Panic, M., Lugonja, P., and Crnojevic, V. (2016). Soybean varieties portfolio optimisation based on yield prediction. *Computers and Electronics in Agriculture*, 127:467–474.

Messina, C., Hammer, G., Dong, Z., Podlich, D., and Cooper, M. (2009). Modelling crop improvement in a g× e× m framework via gene-trait-phenotype relationships. *Crop Physiology: Interfacing with Genetic Improvement and Agronomy. The Netherlands: Elsevier*, pages 235–265.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F., Pérez-Hernández, O., Eskridge, K. M., and Rutkoski, J. (2016). A genomic bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics*, pages g3–116.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Montesinos-López, J. C., Singh, P., Juliana, P., and Salinas-Ruiz, J. (2017). A bayesian poisson-lognormal model for count data for multiple-trait multiple-environment genomic-enabled prediction. *G3: Genes, Genomes, Genetics*, pages g3–117.

Onoufriou, G., Hanheide, M., and Leontidis, G. (2023). Premonition net, a multi-timeline transformer network architecture towards strawberry tabletop yield forecasting. *Computers and Electronics in Agriculture*, 208:107784.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics*, 97(1-2):195–201.

Romero, J. R., Roncallo, P. F., Akkiraju, P. C., Ponzoni, I., Echenique, V. C., and Carballido, J. A. (2013). Using classification algorithms for predicting durum wheat yield in the province of buenos aires. *Computers and Electronics in Agriculture*, 96:173–179.

Sa'diyah, H. and Hadi, A. F. (2016). Ammi model for yield estimation in multi-environment trials: a comparison to blup. *Agriculture and Agricultural Science Procedia*, 9:163–169.

scikit-learn development team (2007-2023a). scikit-learn decisiontreeregressor documentation. Accessed: 01/07/2024.

scikit-learn development team (2007-2023b). scikit-learn kneighborsregressor documentation. Accessed: 01/07/2024.

scikit-learn development team (2007-2023c). scikit-learn lassocv documentation. Accessed: 01/06/2024.

scikit-learn development team (2007-2023d). scikit-learn randomforestregressor documentation. Accessed: 01/07/2024.

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., and Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports*, 12(1):3215.

Team, L. D. (2024). Lightgbm lgbmregressor documentation. Accessed: 01/07/2024.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

CHAPTER 5. GENERAL CONCLUSION

In conclusion, the studies presented in this dissertation demonstrated the importance of advanced optimization and deep learning techniques in improving various aspects of crop production and management. The results suggest that these techniques can provide valuable insights into the complex relationships between genetic and environmental factors and help optimize crop yield and productivity. These findings have important implications for the agricultural industry and can potentially lead to more sustainable and efficient crop production practices in the future.

Addressing the intricate dynamics of corn planting and harvest scheduling, the dissertation first tackled the complexities arising from diverse corn hybrids with distinct planting windows. Leveraging optimization methodologies including mixed-integer linear programming (MILP) models and heuristic algorithms, the first study proposed innovative solutions to optimize planting and harvesting dates considering varying storage capacity scenarios and GDU scenarios. Furthermore, it incorporated deep learning techniques such as recurrent neural networks (RNNs) to predict growing degree units (GDUs), essential for scheduling amidst uncertain growing conditions. The proposed approach demonstrated superior performance providing optimal planting and harvesting schedules for different storage capacity scenarios. This approach has the potential to help year round seed corn producers in achieving consistent weekly harvest quantities while minimizing the logistical and productivity issues associated with planting and harvesting schedules. The main contributions of our first paper are as follow:

Contributions to the Field of Crop Harvest and Planting Scheduling: Application of Optimization for Optimal Planting and Harvesting Dates

- Proposed innovative solutions to optimize planting and harvesting dates for diverse corn hybrids with distinct planting windows.

- Utilized mixed-integer linear programming (MILP) models and heuristic algorithms to address complexities in scheduling under varying storage capacity scenarios.
- Demonstrated the potential of the approach to achieve consistent weekly harvest quantities, thereby minimizing logistical and productivity issues.

Contributions to Optimization Modeling for Crop Planting and Harvest Scheduling: Innovative Optimization Methodology for Enhanced Scheduling Efficiency

- Incorporated deep learning techniques, such as recurrent neural networks (RNNs), to predict growing degree units (GDUs) for scheduling amidst uncertain growing conditions.
- Integrated recurrent neural networks, mixed-integer linear programming models, and heuristic algorithms to develop a comprehensive approach for optimal planting and harvesting schedules.
- Showcased superior performance of the proposed approach in providing optimal schedules for different storage capacity scenarios.

The second paper presented two novel convolutional neural network (CNN) architectures tailored for soybean yield prediction. The first model, CNN-DNN, combined CNN and fully-connected (FC) neural networks, while the second model, CNN-LSTM-DNN, incorporated a long short-term memory (LSTM) layer for weather variables. Leveraging the Generalized Ensemble Method (GEM), we optimized the weights of these models, achieving superior accuracy compared to baseline models. The GEM model exhibited lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), along with higher correlation coefficients, validating its effectiveness in yield prediction. Utilizing the CNN-DNN model, we identified optimal genotypes for diverse locations and weather conditions, facilitating yield predictions for various scenarios. The inclusion of unique genotype information enabled exploration of genotype-based planting strategies, particularly beneficial in scenarios with limited testing years. Furthermore, our feature importance analysis revealed key predictors influencing model predictions, with location, maturity group (MG), year,

and genotype emerging as significant factors. In the weather category, variables like maximum direct normal irradiance (MDNI) and average precipitation (AP) demonstrated notable impact on predictions. Additionally, we investigated the impact of integrating state-level soil data into our models. Despite data constraints limiting the availability of detailed soil information, our findings suggest that the inclusion of soil variables did not significantly enhance predictive capabilities under current conditions. The main contributions of our second paper include the following aspects:

Contributions to Genotype by Environment Selection and Soybean Yield Prediction: Application of Data-Driven Deep Learning-Based Approaches

- Developed deep learning models to predict soybean yield using a combination of factors, including maturity group, genotype ID, year, location, and weather data.
- Demonstrated the effectiveness of the GEM model in soybean yield prediction.
- Employed the CNN-DNN model to identify optimal genotypes for diverse locations and weather conditions, facilitating yield predictions for all potential genotypes in each specific setting.

Contributions to Deep Learning Modeling: Enhancing Genotype by Environment Selection and Crop Yield Prediction through Data-Driven CNN-Based Deep Learning Models

- Introduced two novel CNN architectures incorporating a 1-D convolution operation and an LSTM layer to capture the nonlinear nature of weather data for crop yield prediction and modeling genotype by environment interactions.
- Employed the Generalized Ensemble Method (GEM) to determine optimal weights for the proposed CNN-based models, resulting in superior performance compared to baseline models.
- Addressed the challenge of genotype by environment interaction in crop yield prediction, offering a data-driven paradigm for genotype selection.

- Conducted comprehensive evaluations comparing the proposed GEM model against commonly used prediction models such as RF, XGBoost, and LASSO.
- Conducted a feature importance analysis to identify significant predictors affecting model predictions, with location, maturity group (MG), year, and genotype emerging as crucial variables.
- Investigated temporal aspects of weather data to identify significant time periods influencing soybean growth stages and yield outcomes.
- Explored the impact of soil variables on model performance and highlighted the potential for incorporating location-specific soil attributes to enhance predictive accuracy in regions where soil quality significantly influences agricultural outcomes.

In the last study, we proposed a hybrid approach that combines transformer models with fully connected neural networks to handle sequential data for predicting crop yields. Our model considered various types of data, including weather patterns, APSIM data, and genetic markers, to capture the complex relationships between different variables. We designed four sets of transformer models tailored to different data types and incorporated additional information such as traits and metadata into fully connected neural networks. We systematically evaluated the impact of different variable combinations on prediction accuracy using test data from the year 2021. By merging various datasets and creating eight unique combinations, each corresponding to a specific model configuration, we demonstrated the superior performance of our Transformer-Enhanced models over six baseline machine learning models. Notably, our model showed the best results on VC4, which included weather and genotype data while excluding APSIM and soil data. Furthermore, we compared our Transformer-Enhanced models with one-dimensional convolutional neural networks (CNNs) and found that our proposed approach outperformed CNNs in handling sequential data. This was particularly evident in VC5, which showed significant decreases in RMSE when using our model. Finally, we extended our analysis to include temporal, genomic, and geographic extrapolations to assess the robustness of our model. The results confirmed that our

Transformer-Enhanced models effectively generalized yield predictions to untested years, hybrids, and locations. The primary contributions of our third paper encompass the following aspects:

Contributions to Crop Yield Prediction: Application of Transformer-Based Neural Networks Models

- Proposed a hybrid transformer-fully connected neural networks framework for crop yield prediction, adeptly handling sequential data with temporal dependencies in weather, spatial and temporal correlations in APSIM data, and genetic linkages among adjacent genetic markers.
- Designed four distinct sets of two-layer transformer models for different data types: weather features (W-transformer), APSIM features focusing on EC-phenological period-soil layer (APS-transformer), APSIM features focusing on EC-phenological period (AP-transformer), and genotype features (G-transformer).
- Investigated the impact of various combinations of variables on prediction errors, utilizing a systematic approach to merge diverse datasets including genotype, trait, metadata, soil, weather, and APSIM data.

Contributions to Deep Learning Modeling: Enhancing Crop Yield Prediction through Transformer-Based Neural Networks Models

- Developed eight unique combinations of datasets, each used to develop a specific Transformer-Enhanced Neural Networks model, to evaluate the combined effects of different variables on crop yield prediction.
- Demonstrated superior performance of the proposed Transformer-Enhanced models over six baseline ML models across diverse variable combinations, particularly showcasing effectiveness in handling sequential data.

- Conducted a comparison between Transformer-Enhanced neural networks models and one-dimensional convolutional neural networks (CNNs), highlighting the former's effectiveness in managing sequential data.
- Extended analysis to include temporal, genomic, and geographic extrapolations, showcasing the robustness of the proposed Transformer-Enhanced model across different variable combinations.

In future research, there are several potential avenues for further exploration based on the findings and limitations identified in the current studies. Firstly, there is potential to delve into the socio-economic implications of optimized planting and harvest scheduling strategies. Investigating their impacts on farm profitability, labor requirements, and market competitiveness would offer valuable insights into the broader economic and social dimensions of agricultural production, thereby contributing to the overall sustainability and resilience of farming systems. Moreover, the integration of additional data sources, such as satellite imagery, remote sensing data, or historical crop yield records, presents an opportunity to enhance prediction accuracy and robustness. By leveraging these supplementary sources of information, researchers can refine predictive models and gain deeper insights into crop growth dynamics and environmental influences. Furthermore, exploring ensemble learning techniques to combine predictions from multiple models could be beneficial. By harnessing the complementary strengths of different approaches, ensemble methods have the potential to improve overall prediction performance and enhance the reliability of yield forecasts. Addressing the current limitations in soil data availability is also crucial for future research endeavors. Incorporating more detailed soil information, including soil composition, texture, and fertility levels, could provide a more comprehensive understanding of soil-crop interactions and significantly enhance prediction accuracy. Overall, these potential avenues for further exploration offer exciting opportunities to advance our understanding of agricultural systems and enhance the effectiveness of predictive modeling techniques in agriculture. By addressing these research gaps, future studies can contribute to the development of more accurate, reliable, and actionable insights for agricultural decision-making and management.